UDC 004.891.032.26:629.7.01.066

MODEL-AGNOSTIC META-LEARNING FOR RESILIENCE OPTIMIZATION OF ARTIFICIAL INTELLIGENCE SYSTEM

Moskalenko V. V. – PhD, Associate Professor, Associate Professor of Computer Science department, Sumy State University, Sumy, Ukraine.

ABSTRACT

Context. The problem of optimizing the resilience of artificial intelligence systems to destructive disturbances has not yet been fully solved and is quite relevant for safety-critical applications. The task of optimizing the resilience of an artificial intelligence system to disturbing influences is a high-level task in relation to efficiency optimization, which determines the prospects of using the ideas and methods of meta-learning to solve it. The object of current research is the process of meta-learning aimed at optimizing the resilience of an artificial intelligence system to destructive disturbances. The subjects of the study are architectural add-ons and the meta-learning method which optimize resilience to adversarial attacks, fault injection, and task changes.

Objective. Stated research goal is to develop an effective meta-learning method for optimizing the resilience of an artificial intelligence system to destructive disturbances.

Method. The resilience optimization is implemented by combining the ideas and methods of adversarial learning, fault-tolerant learning, model-agnostic meta-learning, few-shot learning, gradient optimization methods, and probabilistic gradient approximation strategies. The choice of architectural add-ons is based on parameter-efficient knowledge transfer designed to save resources and avoid the problem of catastrophic forgetting.

Results. A model-agnostic meta-learning method for optimizing the resilience of artificial intelligence systems based on gradient meta-updates or meta-updates using an evolutionary strategy has been developed. This method involves the use of tuner and meta-tuner blocks that perform parallel correction of the building blocks of a original deep neural network. The ability of the proposed approach to increase the efficiency of perturbation absorption and increase the integral resilience indicator of the artificial intelligence system is experimentally tested on the example of the image classification task. The experiments were conducted on a model with the ResNet-18 architecture, with an add-on in the form of tuners and meta-tuners with the Conv-Adapter architecture. In this case, CIFAR-10 is used as a base set on which the model was trained, and CIFAR-100 is used as a set for generating samples on which adaptation is performed using a few-shot learning scenarios. We compare the resilience of the artificial intelligence system after pre-training tuners and meta-tuners using the adversarial learning algorithm, the fault-tolerant learning algorithm, the conventional model-agnostic meta-learning algorithm, and the proposed meta-learning method for optimizing resilience. Also, the meta-learning algorithms with meta-gradient updating and meta-updating based on the evolutionary strategy are compared on the basis of the integral resilience indicator.

Conclusions. It has been experimentally confirmed that the proposed method provides a better resilience to random bit-flip injection compared to fault injection training by an average of 5%. Also, the proposed method provides a better resilience to L_∞-adversarial evasion attacks compared to adversarial training by an average of 4.8%. In addition, an average 4.8% increase in the resilience to task changes is demonstrated compared to conventional fine-tuning of tuners. Moreover, meta-learning with an evolutionary strategy provides, on average, higher values of the resilience indicator. On the downside, this meta-learning method requires more iterations.

KEYWORDS: Meta-learning, Evolutionary Strategies, Parameter-Efficient Transfer Learning, Robustness, Resilience, Adversarial Attacks, Faults Injection, Few-Shot Learning.

ABBREVIATIONS

AIS is an Artificial Intelligence System;

CIFAR is a Canadian Institute for Advanced Research dataset;

CMA-ES is the covariance matrix adaptation evolution strategy optimization algorithm;

MAML is a Model-Agnostic Meta-Learning;

PID is a proportional-integral-derivative controller.

NOMENCLATURE

$\overline{Acc}$ is an accuracy averaged over set of disturbance implementations;

α is a step size hyperparameter for inner loop of meta-learning;

β is a step size hyperparameters for outer loop of meta-learning;

γ is a tuner' hyperparameter which regulates channel compression by 1, 2, 4, or 8 times;

D is a dataset for sampling few-shot learning tasks;

D_{base} is a dataset for main task;

D_{base}^{tr} is a training subset for main task;

D_{base}^{val} is an evaluation subset for main task;

D_k^{tr} is a training subset for k-th few-shot learning task;

D_k^{val} is an evaluation subset for k-th few-shot learning task;

F is an expected value of resilience criterion;

ϕ is a set of parameters of tuners;

ϕ^* is a set of optimal parameters of tuners;

g is a perturbation vector formed for the parameters that being meta-optimized;

K is a number of few-shot learning tasks;

DOI 10.15588/1607-3274-2023-2-9

k_{shot} is a number of images per class in few-shot task;

L_0-norm is a count the number of non-zero elements in an adversarial perturbation vector;

L_∞-norm is a largest magnitude among each element of an adversarial perturbation vector;

L_{τ_i} is a loss function for current task;

N is a number of disturbance implementations;

n is a number of implementations of disturbances of the same type on one meta iteration;

n_{way} is a number of classes in few-shot task;

P_0 is a model performance before disturbance impact;

$p(\tau)$ is a distribution of disturbance implementations;

P_{τ_i} is a performance metric for current state of model parameters and evaluation data;

$\overline{R}$ is an integral resilience indicator averaged over set of disturbance implementations;

R_{τ_i} is a function that calculates the value of the integral resilience indicator for τ_i disturbance implementation;

σ is a precision parameter for evolution strategies of meta-learning;

θ are parameters of pretrained and frozen base AIS model;

Ξ is a set of all parameters of AIS model;

T is a maximum number of adaptation steps;

τ_i is an i-th implementation of disturbance;

U_{τ_i} is an operator that combines disturbance generation and adaptation;

ω is a set of parameters of meta-tuners;

ω^* is a set of optimal parameters of meta-tuners;

W is a task specified parameters;

W_{base} is a head weights for the main task.

INTRODUCTION

AIS are vulnerable to various types of disturbances. The most studied types of disturbances are fault injection, adversarial attacks, drift, and out-of-distribution data [1]. These disturbances can lead to financial and human life losses in safety-critical applications. This amplifies the need for research on vulnerabilities and resilience aspects of AISs.

There are many papers devoted to protecting AIS against fault injection, adversarial attacks, various types of drift, and out-of-distribution [2, 3, 4]. However, few studies have investigated the compatibility of protection mechanisms against different types of disturbances. There is a lack of research on simultaneous protection against the impact of different types of disturbing factors.

Since the main goal of machine learning of AIS is to ensure their maximum performance, the task of increasing the resilience of an AIS to disturbing influences can be considered an additional, higher-level task. In this case, the use of ideas and methods of meta-learning to ensure the resilience of an artificial intelligence system seems to be the most reasonable and promising approach. In addition, different mechanisms for protecting against disturbances may be incompatible or unbalanced, and meta-learning can be seen as a way to harmonize the effect of different machine learning mechanisms to enhance a high-level goal of resilience.

The concept of resilience includes both architectural aspects related to system redundancy and mechanisms of graceful degradation, and behavioral aspects related to the ability to absorb disturbances (robustness) and the ability to quickly adapt and evolve [5, 6]. All these aspects are interconnected in neural networks, especially in the context of defense against various types of disturbing factors. However, conventional meta-learning methods usually take into account only one aspect of resilience and only one type of disturbance.

Conventional methods for optimizing AIS parameters typically aim to maximize the performance metric, sometimes with simultaneous improvements in robustness [7]. Meta-learning algorithms usually solve the problem of learning in a few shots and the problem of convergence in a minimum number of iterations [8]. However, there is a lack of research on meta-learning application to simultaneously optimize different components of resilience, especially in the context of different types of disturbances.

As the size of neural networks grows, more and more attention is focused on parameter-efficient fine-tuning methods for rapid adaptation to new tasks and domains [9]. This approach does not change the original AIS model, but instead adds additional elements that are fine-tuned to improve the resulting model. Parameter-efficient fine-tuning is also relevant for adapting to disruptive influences in the context of affordable resilience, as there are always resource constraints in practice.

The object of research is the process of meta-learning for artificial intelligence system which functions under influences of destructive perturbations and is subject to resource constraints.

The subjects of the research are model-agnostic method of meta-learning of an artificial intelligence system that provide resilience to adversarial attacks, fault injection attacks and task changes.

The goal is a development parameter-efficient model-agnostic meta-learning method of artificial intelligence system which provides resilience to adversarial attacks, fault injections and task changes.

1 PROBLEM STATEMENT

Let $\{\tau_i | i=\overline{1,N}\}$ is set of disturbance implementations for AIS. Disturbances τ_i can be considered as adversarial attacks, fault injection, or switching to a new task. Let $\{D_{base}=\{D_{base}^{tr}; D_{base}^{val}\}$ is a dataset on which the model was trained to perform the main task under known conditions. It is also given a dataset

DOI 10.15588/1607-3274-2023-2-9

$D=\{D_k^{tr}; D_k^{val} | k=\overline{1,K}\}$ for K few-shot learning tasks, where fine-tuning data D_k^{tr} is used in the fine-tuning stage and validation set D_k^{val} is used in the meta-update stage. There is also a given set of parameters θ, ϕ, ω and W, where θ are parameters of a pretrained and frozen base AIS model, ϕ and ω are adaptation parameters of AIS model backbone, and W are task specified parameters (model head parameters). Head weights W_{base} for the main task are pre-trained on the data D_{base}.

It is necessary to find such values of the parameters ω^*, ϕ^* which ensure the maximum expected resilience of AIS to the impact of various types of disturbances

$$\max_{\omega,\phi} \underset{\tau_i \sim p(\tau)}{E} [R_{\tau_i}(U_{\tau_i}(\theta,\phi,\omega,W,D))]. \quad (1)$$

The operator U should combine a disturbing influence and adaptation in T steps, which maps the current state of ϕ to new state of ϕ. Adversarial attacks, fault injection, or switching to new tasks may be considered as i-th disturbance τ_i

The function R_{τ_i} calculates the value of the integral resilience indicator for a particular disturbance implementation over model parameters ω during its adaptation and the test sample $D_{\tau_i}^{val}$. R_{τ_i} is a function of the performance metric P_{τ_i}.

$$R_{\tau_i} = \frac{1}{P_0 T} \sum_{t=1}^{T} P_{\tau_i}(\theta,\omega,\phi_t,W_t,D_{\tau_i}^{val}). \quad (2)$$

Gradient-based meta-learning requires finding such values of the parameters ω^*, ϕ^* that will ensure the minimum expected loss function L on the set of implementations of different types of disturbances τ_i

$$\min_{\omega,\phi} \underset{\tau_i \sim p(\tau)}{E} [L_{\tau_i}(U_{\tau_i}(\theta,\phi,\omega,W,D))]. \quad (3)$$

2 REVIEW OF THE LITERATURE

The problem of the resilience of AIS to various types of disturbances was highlighted in [1, 5, 6]. In [10, 11] it is noted that the two main characteristics of the behavior of a resilient system are the simultaneous absorption of disturbances (robustness) and rapid adaptation to new disturbances. However, the majority of machine learning methods focus on maximizing pefomance under normal conditions, and sometimes one aspect of resilience [12, 13].

Increasing the robustness of AIS to adversarial attacks is based on gradient masking [14], robustness optimization [15], and adversarial attack detection [16]. Fault tolerance is implemented through the application of fault masking methods [17], explicit redundancy input methods [18], and error detection methods [19]. The most popular method is adversarial training, which is related to robustness optimization. It consists in generating perturbations of training samples during training. Increasing the robustness to small domain shifts or task changes is provided by Out-of-domain generalization methods [20]. However, there is a lack of methods that ensure optimal robustness to the complex effects of various types of disturbances.

Increasing the speed of adaptation is usually implemented by improving optimizers or by reducing the requirements for the amount of training data. Research [21, 22] considers the use of PID control principles that can increase the stability and speed of the process of optimizing system parameters. In [23], methods for training neural network optimizers to increase the learning speed are considered.

One of the ways to increase the generalization capability of the network and reduce the requirements for training data while adapting to changes is meta-learning. In [24], various model-agnostic meta-learning methods for implementing Few-Shot Learning are considered. In [7], an attempt was made to integrate different methods for increasing the robustness of a neural network with meta-learning for Few-Shot Learning. As a result, it has been demonstrated that the incorporation of regularization and the introduction of perturbative effects can be effectively executed in both the internal loop and the external loop of meta-learning. Another study [25] demonstrated that the outer loop of meta-learning could be implemented using an evolutionary strategy, enabling the use of even non-differentiable, non-smooth, and non-decomposable meta-objectives. However, there is a lack of research examining meta-objectives such as robustness, adaptation speed, or integral resilience indicators.

Study [26], has shown that ensuring resilience may require significant resources, and it is advisable to consider approaches for providing affordable resilience. In the works [9], it is proposed to perform model adaptation for specific tasks or domains within the framework of parameter-efficient transfer learning. In this approach, the large AIS model remains unchanged (frozen). Instead, the frozen model is modulated by adding adapters with a small number of parameters. In the study [27], the use of adapters in conjunction with meta-adapters was considered to further enhance adaptation efficiency. However, the possibility of using adapters for optimizing model robustness and adaptation speed to new types of disturbances has not yet been explored.

Thus, there is a lack of research on the use of meta-learning for ensuring the resilience of AIS to the combined impact of various types of disturbances. However, existing studies demonstrate the possibility of optimizing individual aspects of resilience. Therefore, it is relevant to investigate the potential for combining different mechanisms to ensure the resilience of AIS in the full sense of this concept.

DOI 10.15588/1607-3274-2023-2-9

3 MATERIALS AND METHODS

The larger the model, the more computationally complex it is to fine-tune for adaptation to new conditions. Moreover, there is a potential risk of catastrophic forgetting under the influence of new information. Therefore, it is proposed to attach tuners to the model, which can be computationally efficient in fine-tuning [9]. In this case, the weight coefficients of the model remain frozen. The original model usually consists of certain blocks or modules, such as Convolutional Residual Block, Multi-layer Perceptron Block, Multi-head Self-Attention Block, and others. To generalize, we will refer to these blocks as frozen operations and denote them as OP(x). The parallel method of connecting a tuner (adapter) to the frozen blocks of the model is the most convenient and versatile approach (Fig. 1a) [28]. In this case, to ensure the properties of resilience, it is proposed to use three consecutive blocks of tuners at once, two of which are tuned during meta-training (Fig. 1a) [27]. To balance between different modules, we introduce a channel-wise scaling factor.

For the same frozen model block, the architecture of the tuners is chosen to be identical. Various tuner (adapter) architectures have become popular in the literature, with the most computationally efficient ones shown in Fig. 1b, Fig. 1c, and Fig. 1d. The architectures depicted in Fig. 1b and Fig. 1c are based on convolutional layers, making them computationally simpler on the one hand and characterized by lower capacity for absorbing disturbances on the other hand. However, convolutional tuners have proven themselves effective in correcting frozen convolutional blocks [29]. The convolutional tuner shown in Fig. 1c has a hyperparameter γ, which regulates channel compression by 1, 2, 4, or 8 times.

If we reject the specialization of different parameters of the AI model and denote the set of all parameters as $\Xi =< \theta, \phi, \omega, W >$, then the process of meta-learning using gradient learning methods can be described by the formula

$$\Xi^* = \arg\min_{\Xi} \underset{\tau_i \sim p(\tau)}{E} [L_{\tau_i}(U_{\tau_i}(\Xi, D_{\tau_i}))]. \quad (4)$$

If we use the SGD stochastic gradient descent algorithm with T steps in the U operator and use gradient meta-update in the outer loop, we will get the algorithm shown in Fig. 2. Moreover, to simplify computations and increase stability, the tunable and meta-tunable parameters can be updated using the REPTIL algorithm [30].

If we do not restrict ourselves to gradient algorithms in the outer loop, then the meta-learning for direct maximization of the expected resilience criterion can be described by the formula

$$\Xi^* = \arg\max_{\Xi} \underset{\tau_i \sim p(\tau)}{E} [R_{\tau_i}(U(\Xi, D))] = \\ = \arg\max_{\Xi} F(\Xi). \quad (5)$$

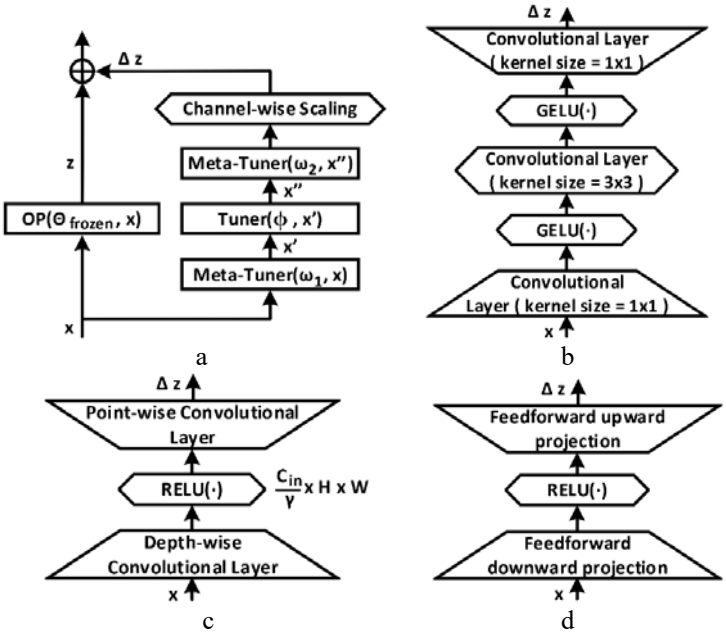

Figure 1 – Parallel tunning scheme and tuner architectures: a – parallel tuning scheme for the frozen block; b – tuner or meta-tuner based on ResNet-like convpass; c – tuner or meta-tuner based on two-layer convolutional network with channel dimension down-sampling bottleneck; d – tuner or meta-tuner based on two-layer feed-forward network with a downward projection bottleneck

DOI 10.15588/1607-3274-2023-2-9

Require: Distribution over disturbances $p(\tau)$; Step size hyperparameters α, β;
Number of adaptation steps T.

Pretrain ϕ, ω on original data D_{base}
While not done do:
  Select type of disturbance from set {fault injection, evasion adversarial attack, task change}
  Sample disturbance implementations $\tau_i \sim p(\tau)$, $i = \overline{1,n}$
  For i=1,2,…,n do:
    Clone the current parameters: $\dot{\theta}_{\tau_i}, \acute{\omega}_{\tau_i}, \acute{\phi}_{\tau_i}, \acute{W}_{\tau_i} \leftarrow copy(\theta, \omega, \phi, W_{base})$
    If disturbance type is a task change:
      Sample the training and validation data $D_{\tau_i}^{tr}$, $D_{\tau_i}^{val}$ from new task
    else:
      Sample the training and validation data $D_{\tau_i}^{tr}$, $D_{\tau_i}^{val}$ from D_{base}
    If disturbance type is a fault injection:
      $\dot{\theta}_{\tau_i}, \acute{\omega}_{\tau_i}, \acute{\phi}_{\tau_i} \leftarrow Fault\_injection(\dot{\theta}_{\tau_i}, \acute{\omega}_{\tau_i}, \acute{\phi}_{\tau_i}, \acute{W}_{\tau_i})$
    If disturbance type is an evasion adversarial attack:
      $D_{\tau_i}^{tr}, D_{\tau_i}^{val} \leftarrow Adversarial\_perturbation(D_{\tau_i}^{tr}, D_{\tau_i}^{val})$
    $\phi_{\tau_i} \leftarrow SGD_{\phi,W}(L_{\tau_i}(\dot{\theta}_{\tau_i}, \acute{\omega}_{\tau_i}, \acute{\phi}_{\tau_i}, \acute{W}_{\tau_i}, D_{\tau_i}^{tr}), T, \alpha)$
  $\omega \leftarrow \omega - \beta \nabla_{\omega} \sum_{\tau_i \sim p(\tau)} L_{\tau_i}(\theta, \omega, \phi_{\tau_i}, D_{\tau_i}^{val})$
  $\phi \leftarrow \phi + \beta \frac{1}{n} \sum_{i=1}^{n} (\phi_{\tau_i} - \phi)$

Figure 2 – Pseudocode of model-agnostic gradient-based meta-learning for AIS resilience optimization

Require: Distribution over disturbances $p(\tau)$; Step size hyperparameters α, β;
Precision parameter σ; Number of adaptation steps T.

Pretrain ϕ, ω on original data D_{base}
While not done do:
  Select type of disturbance from set { fault injection, evasion adversarial attack, task change}
  Sample disturbance implementations $\tau_i \sim p(\tau)$, $i = \overline{1,n}$
  Sample perturbation vectors $g_{\phi,\tau_i} \sim N(0, I)$, $i = \overline{1,n}$; $g_{\omega,\tau_i} \sim N(0, I)$, $i = \overline{1,n}$
  For i=1,2,…,n do:
    Clone the current parameters: $\dot{\theta}_{\tau_i}, \acute{\omega}_{\tau_i}, \acute{\phi}_{\tau_i}, \acute{W}_{\tau_i} \leftarrow copy(\theta, \omega, \phi, W_{base})$
    $\acute{\phi}_{\tau_i+} \leftarrow \acute{\phi}_{\tau_i} + \sigma g_{\phi,\tau_i}$; $\acute{\phi}_{\tau_i-} \leftarrow \acute{\phi}_{\tau_i} - \sigma g_{\phi,\tau_i}$
    $\acute{\omega}_{\tau_i+} \leftarrow \acute{\omega}_{\tau_i} + \sigma g_{\omega,\tau_i}$; $\acute{\omega}_{\tau_i-} \leftarrow \acute{\omega}_{\tau_i} - \sigma g_{\omega,\tau_i}$
    If disturbance type is a task change:
      Sample the training and validation data $D_{\tau_i}^{tr}$, $D_{\tau_i}^{val}$ from new task
    else:
      Sample the training and validation data $D_{\tau_i}^{tr}$, $D_{\tau_i}^{val}$ from D_{base}
    If disturbance type is a fault injection:
      $\dot{\theta}_{\tau_i}, \acute{\omega}_{\tau_i+}, \acute{\omega}_{\tau_i-}, \acute{\phi}_{\tau_i+}, \acute{\phi}_{\tau_i-} \leftarrow Fault\_injection(\dot{\theta}_{\tau_i}, \acute{\omega}_{\tau_i+}, \acute{\omega}_{\tau_i-}, \acute{\phi}_{\tau_i+}, \acute{\phi}_{\tau_i-})$
    If disturbance type is an evasion adversarial attack:
      $D_{\tau_i}^{tr}, D_{\tau_i}^{val} \leftarrow Adversarial\_perturbation(D_{\tau_i}^{tr}, D_{\tau_i}^{val})$
    $\{\acute{\phi}_{\tau_i+,t} | t = \overline{1,T}\} \leftarrow SGD_{\phi,W}(L_{\tau_i}(\dot{\theta}_{\tau_i}, \acute{\omega}_{\tau_i+}, \acute{\phi}_{\tau_i+}, \acute{W}_{\tau_i}, D_{\tau_i}^{tr}), T, \alpha)$
    $\{\acute{\phi}_{\tau_i-,t} | t = \overline{1,T}\} \leftarrow SGD_{\phi,W}(L_{\tau_i}(\dot{\theta}_{\tau_i}, \acute{\omega}_{\tau_i-}, \acute{\phi}_{\tau_i-}, \acute{W}_{\tau_i}, D_{\tau_i}^{tr}), T, \alpha)$
    $R_{+,\tau_i} \leftarrow R(\{P_t(\dot{\theta}_{\tau_i}, \acute{\omega}_{\tau_i+}, \acute{\phi}_{\tau_i+,t}, D_{\tau_i}^{val})\})$
    $R_{-,\tau_i} \leftarrow R(\{P_t(\dot{\theta}_{\tau_i}, \acute{\omega}_{\tau_i-}, \acute{\phi}_{\tau_i-,t}, D_{\tau_i}^{val})\})$
    $R_{\tau_i} \leftarrow \frac{1}{2}(R_{+,\tau_i} - R_{-,\tau_i})$
  $\omega \leftarrow \omega + \beta \frac{1}{\sigma n} \sum_{i=1}^{n} R_{\tau_i} g_{\omega,\tau_i}$
  $\phi \leftarrow \phi + \beta \frac{1}{\sigma n} \sum_{i=1}^{n} R_{\tau_i} g_{\phi,\tau_i}$

Figure 3 – Pseudocode of model-agnostic meta-learning with evolution strategies for AIS resilience optimization

DOI 10.15588/1607-3274-2023-2-9

It is proposed to use an evolutionary strategy for the direct maximization of the resilience criterion due to the possible non-differentiability of the criterion. In this case, the gradient estimation can be performed over the Gaussian-smoothed version of the outer loop objective, which is calculated by the formula [8]

$$\nabla \underset{g\sim N(0,I)}{E}[F(\Xi+\sigma g)] = \\ = \frac{1}{2\sigma} E[R(\Xi+\sigma g) - R(\Xi-\sigma g)]. \quad (6)$$

If, in the meta-optimization algorithm with an evolutionary strategy, a single perturbation vector g is formed for the meta-optimized parameters at the beginning of each meta-optimization iteration, the resulting algorithm will be as shown in Fig. 3.

The analysis of Fig. 2 shows that the type of disruptive influence does not change within a single meta-adaptation step. However, each meta-adaptation step begins with the selection of a disruptive influence type, followed by the generation of n implementations of the disruptive influence with a subsequent nested adaptation loop for each of them. Simultaneously combining disturbances may be ineffective. For example, after adding fault injection to the weights, we will have an outdated model, and applying adversarial attacks to it may be irrelevant.

The formation of adversarial samples is based on the $Adversarial\_perturbation()$ function. For differentiated models, FGSM attacks or PGD attacks can be used [31]. It is proposed to use adversarial attacks based on the search algorithm of the covariance matrix adaptation evolution strategy (CMA-ES) for non-differentiable models [32]. The level of perturbation is limited by the L_∞ -norm or L_0 -norm. In this case, if the image is normalized by dividing pixel brightness by 255, then the specified disturbance level is also divided by 255.

The formation of fault injections is performed by the $Fault\_injection()$ [33]. It is suggested to choose the most difficult fault type to absorb, which involves generating an inversion of a randomly selected bit (bit-flip injection) in the weight coefficient of the model. For non-differentiable models, it is proposed to generate up to 100 damaged weight versions, selecting the one providing the greatest reduction in accuracy. For differentiable models, it is suggested to pass the test dataset through the network and calculate the gradients, which can then be sorted by their absolute values. In the top-k weights with the highest gradient, one bit is inverted in a random position. The proportion of weights for which one random bit is inverted can be denoted as the fault rate.

Task change is needed to simulate concept drift and out-of-distribution. Forming a sample of other tasks can be done by randomizing the domain of the same task or by selecting tasks from relevant domains but sampling truly different tasks. These two approaches can also be combined. In any case, an attempt should be made to sample data from larger and more diverse sets than D_{base}.

The analysis of Fig. 3 shows that, to calculate the resilience indicator, it is necessary to calculate the intermediate values of the performance metric for T adaptation steps, which in most cases is not differentiated.

4 EXPERIMENTS

For simplicity of experiments, we will use the Resnet-18 as the base model, pretrained on CIFAR-10 [34]. The architecture of the tuner and meta-tuner is chosen to be the same in the form of a two-layer convolutional network with channel dimension down-sampling bottleneck $(\gamma = 2)$. To generate new tasks, CIFAR-100 is used, from which data is sampled for randomly selected 10 classes $(n_{way} = 10)$ [34]. It is proposed to use 16 images per class $(k_{shot} = 16)$, which are sent in mini-batches of 4 images $(mini\_batch\_size = 4)$ during adaptation. Thus, the number of adaptation steps is $T = (k_{shot} * n_{way}) / mini\_batch\_size = 40$ iterations. The base task is used during meta-learning along with new tasks. The learning rate of the inner and outer loop of meta-learning are $\alpha = 0.001$ and $\beta = 0.0001$, respectively. The maximum number of meta-iterations is 300. However, the Early Stopping algorithm is used to stop meta-learning, which terminates the execution if the criterion does not change for more than 10 consecutive iterations by more than 0.001.

During the experiments, it is necessary to determine whether the connection of a meta-trained tuners can improve the ability to absorb disturbances such as faults and adversarial attacks, even if they are formed in a way different from what is used during training. It is also planned to determine whether meta-learning improves the speed of adaptation to faults and adversarial attacks compared to training under the influence of disturbances without meta-learning.

When calculating the integral indicator of resilience to disturbance (2), it is proposed to use the accuracy metric as the performance criterion P_{τ_i}. It is necessary to determine the advantages and disadvantages of the gradient-based meta-adaptation algorithm and the meta-adaptation algorithm with an evolutionary strategy.

The experimental research is proposed to be carried out in the following sequence:

1. Testing the pre-trained model without tuners and meta-tuners on disturbances of varying intensity.

2. Adding tuners and meta-tuners trained on data during fault injection into the entire model with a fixed fault_rate for testing the resulting model on different fault_rate values.

3. Adding tuners and meta-tuners trained on data with adversarial disturbances with a fixed perturbation level

DOI 10.15588/1607-3274-2023-2-9

for L_∞ for testing the resulting model on different perturbation levels for L_∞.

4. Adding tuners and meta-tuners trained using the proposed gradient-based meta-learning algorithm for optimizing resilience (Fig. 2) followed by testing of the resulting model on disturbances of varying amplitude.

5. Adding tuners and meta-tuners trained using the proposed meta-learning algorithm with an evolutionary strategy for optimizing resilience (Fig. 3) followed by testing of the resulting model on disturbances of varying level.

6. Adding tuners and meta-tuners and calculating the average resilience value during adaptation to new tasks sampled from CIFAR-100.

7. Adding tuners and meta-tuners trained using the proposed gradient-based meta-learning algorithm for optimizing resilience (Fig. 2) and calculating the average resilience value during adaptation to new tasks sampled from CIFAR-100 but not used during meta-learning.

During the studies of the pre-trained model without tuners, only the average accuracy value on the test set is calculated. During testing of the pre-trained model with tuners and meta-tuners, the meta-tuners remain fixed, and the tuner is used to restore performance and calculate the average resilience value (2).

For training tuners with meta-tuners, fault injection is carried out by selecting weights with the largest absolute gradient values. The proportion of modified weights is $fault\_rate = 0.3$. For testing the resulting model, fault injection will be performed by random bit-flips in randomly selected weights, the proportion of which (fault_rate) ranges from 0.1 to 0.6.

The training of the tuners and meta-tuners involves generating adversarial samples using the FGSM algorithm with perturbation_level according to L_∞ up to 3. However, to test the resulting model against adversarial attacks, the adversarial samples are generated using the CMA-ES algorithm with perturbation_level according to L_∞-norm from 1 to 10.

Taking into account the elements of randomization, it is proposed to use their average values when assessing the accuracy and resilience of the model. To this end, 100 implementations of a certain type of disturbance are generated and applied to the same model or data. The average value of resilience during adaptation to new tasks is estimated on 5 task implementations (combinations of classes from CIFAR-100 that did not participate in the training of the meta-tuners).

5 RESULTS

The results of testing the impact of fault injection on model performance are shown in Table 1. When adding tuners and meta-tuners, before testing, they are pre-trained using gradient-based algorithms with and without disturbances until the base model's accuracy is reached. The average accuracy $\overline{Acc}$ and the integral resilience indicator $\overline{R}$ for the model are evaluated on 100 implementations of fault injection with a given fault rate. During testing, meta-tuners are fixed, and tuners can be used for adaptation to disturbances and calculation of the resilience indicator. Meta-learning is performed using a gradient-based algorithm, where the early stopping condition occurred at the 150-th iteration (Fig. 2).

The analysis of Table 1 shows that tuners with meta-tuners can increase the robustness of the pre-trained model to faults by absorbing the impact. Moreover, the proposed resilience-aware meta-learning method provides a better resilience indicator compared to fault-tolerance training, on average by 5%. This means that during 40 iterations of tuning, performance recovery occurs faster on average if the model was prepared based on resilience-aware meta-learning. The resilience measurements in the experiment have standard deviation not exceeding 1.0.

Table 1 – Experimental data of model resilience to the faults injection testing

Fault rate	Only Pre-trained model	Pretrained model with tuners and meta-tuners trained under fault injection		Pretrained model with tuners and meta-tuners meta-trained for resilience optimization	
	$\overline{Acc}$	$\overline{Acc}$	$\overline{R}$	$\overline{Acc}$	$\overline{R}$
0.0	92.5%	92.8%	–	93.1%	–
0.1	90.2%	91.1%	0.971	92.2%	0.986
0.3	85.1%	87.6%	0.944	89.1%	0.971
0.5	83.0%	85.5%	0.883	86.5%	0.955
0.6	75.4%	80.1%	0.831	84.9%	0.917

The results of testing the impact of adversarial attacks on the model's performance are shown in Table 2. Tuners and meta-tuners are trained without and with perturbations until the resulting model reaches the accuracy of the base model for the next resiliency test. The average value of the accuracy $\overline{Acc}$ and the integral indicator of resilience $\overline{R}$ for the model is estimated on 100 implementations of adversarial perturbations of the dataset with a given perturbation level. After freezing the parameters of the meta-tuners, the tuners can be used to adapt to the disturbance and calculate the resilience indicator. Meta-learning is performed using a gradient-based algorithm (Fig. 2).

Table 2 – Experimental data of model resilience to the adversarial attack testing

Per-tur-ba-tion level	Only Pre-trained model	Pretrained model with tuners and meta-tuners trained under adversarial attack		Pretrained model with tuners and meta-tuners meta-trained for resilience optimization	
	$\overline{Acc}$	$\overline{Acc}$	$\overline{R}$	$\overline{Acc}$	$\overline{R}$
0	92.5%	92.8%	–	92.7%	–
1	91.6%	91.1	0.965	92.0%	0.981
3	88.1%	88.9	0.934	90.1%	0.980
5	82.5%	82.7	0.865	84.8%	0.922
10	74.8%	75.9	0.821	77.7%	0.897

DOI 10.15588/1607-3274-2023-2-9

The analysis of Table 2 shows that meta-tuners can increase the robustness of a trained model to adversarial attacks by absorbing part of the disturbance. Moreover, the proposed method of resilience-aware meta-learning provides a better resilience indicator compared to adversarial training by an average of 4.8%. That is, within 40 iterations of tuning, performance recovery is faster after resilience-aware meta-learning. The resilience measurements in the experiment have standard deviation not exceeding 1.1.

The results of resilience testing of the model with tuners and meta-tuners that were meta-trained with the evolutionary strategy (Fig. 3) are shown in Table 3 and Table 4. In this case, the meta-learning was performed with 233 iterations until the Early Stopping condition was reached.

The analysis of Table 3 and Table 4 shows that meta-learning with an evolutionary strategy also improves the perturbation absorption and the integral resilience indicator. A comparison of the results from Tables 1 and 2 with the results from Tables 3 and 4 shows that the results are comparable, but with a slight advantage in resilience (more than 1.5%) for the evolutionary optimization strategy. It is also worth noting that the evolutionary strategy required 83 additional iterations to achieve the optimal result.

Table 3 – Experimental data of model resilience to the faults injection testing after resilience aware meta-learning with evolution strategies

Fault rate	$\overline{Acc}$	$\overline{R}$
0.0	93.1%	–
0.1	93.3%	0.988
0.3	89.9%	0.979
0.5	87.5%	0.971
0.6	85.0%	0.941

Table 4 – Experimental data of model resilience to the adversarial attack testing after resilience aware meta-learning with evolution strategies

Perturbation level	$\overline{Acc}$	$\overline{R}$
0	92.7%	–
1	93.5%	0.986
3	91.6%	0.984
5	86.7%	0.954
10	83.5%	0.919

The advantage of using meta-learning instead of conventional fine-tuning of tuners was evaluated based on the experiment results which are shown in Table 5.

Table 5 – Experimental data of model resilience to the task change testing

Pretraining method of tuners and meta-tuners	$\overline{R}$
Pre-trained on the base dataset until the accuracy of the base model is achieved	0.933
Meta-trained for resilience optimization	0.981

The analysis of Table 5 shows that when adapting to new tasks, the meta-trained tuners and meta-tuners provide on average a 4.8% higher value of the integrated resilience indicator over 40 iterations of adaptation than a simple pre-training on the base task.

Thus, the proposed meta-learning algorithm for optimizing the AIS resilience ensures an increase in the AIS model's resilience to disturbances compared to conventional approaches such as fault-tolerant training, adversarial training, and fine-tuning.

6 DISCUSSION

Experimental data confirm the increase in the resilience of the AI system to disturbing influences on the example of image classification with the use of the proposed meta-learning method. However, it is not clear exactly what effect the use of different types of perturbations in the internal optimization loop would have on the resilience. It is not known whether the conventional Model-Agnostic Meta-Learning (MAML) algorithm for Few Shot Learning will be inferior to the proposed method. Therefore, it is proposed to compare the results of testing meta-trained tuners and meta-tuners using the conventional MAML and the proposed algorithm (Fig. 2). Table 6 and Table 7 show the results of testing meta-trained tuners with meta-tuners using conventional MAML and the proposed Resilient-aware MAML. In this case, Table 6 illustrates the result of testing for fault injection, and Table 7 illustrates the result of testing for adversarial evasion attacks.

Table 6 – Experimental data of model resilience to the faults injection testing for Conventional MAML and Proposed Resilient-aware MAML

Fault rate	Gradient-based Conventional MAML		Proposed Gradient-based Resilient-aware MAML	
	$\overline{Acc}$	$\overline{R}$	$\overline{Acc}$	$\overline{R}$
0.0	92.8%	–	93.1%	–
0.1	90.7%	0.962	92.2%	0.986
0.3	84.9%	0.938	89.1%	0.971
0.5	82.9%	0.877	86.5%	0.955
0.6	75.6%	0.838	84.9%	0.917

Table 7 – Experimental data of model resilience to the adversarial attack testing for Conventional MAML and Proposed Resilient-aware MAML

Perturbation level	Gradient-based Conventional MAML		Proposed Gradient-based Resilient-aware MAML	
	$\overline{Acc}$	$\overline{R}$	$\overline{Acc}$	$\overline{R}$
0	92.5%	-	92.7%	-
1	91.3%	0.961	92.0%	0.981
3	88.6%	0.921	90.1%	0.980
5	81.9%	0.847	84.8%	0.922
10	73.7%	0.811	80.7%	0.907

The analysis of Table 6 shows that tuners and meta-tuners trained with the conventional MAML algorithm for few-shot learning are inferior on average by more than 5% to the proposed algorithm in terms of resilience. Moreover, a comparison with Table 1 shows that the results of conventional MAML are on average 0.3% lower in terms of resilience than fault-tolerant training algorithms.

The analysis of Table 7 shows that tuners and meta-tuners trained with the conventional MAML algorithm for

DOI 10.15588/1607-3274-2023-2-9

few-shot learning are inferior on average by more than 6% to the proposed algorithm in terms of resilience. Moreover, a comparison with Table 2 shows that the results of conventional MAML are inferior in terms of resilience on average by more than 1% to the adversarial training algorithm.

Thus, the proposed Resilient-aware MAML for model pre-training with tuners and meta-tuners ensures better absorption of fault injection and adversarial attacks and faster adaptation to them. Moreover, the proposed approach provides a higher resilience indicator compared to training separately under the influence of each type of perturbation or on the basis of conventional MAML.

CONCLUSIONS

The **scientific novelty** of the obtained result is the new MAML method for optimizing resilience to fault injection, adversarial attacks, and task change is developed. The method involves the use of tuners and meta-tuners which perform parallel correction of the building blocks of the deep neural network. The proposed meta-learning method consists of generating n implementations of a certain type of disturbance at each iteration of meta-optimization and using the results of adaptation for meta-updating tuners and meta-tuners. In this case, meta-updates can be calculated based on gradients or on an evolutionary strategy

It is experimentally proven that the Proposed Resilient-aware MAML improves the ability of the basic model to absorb disturbances and increases the speed of adaptation compared to conventional approaches. The proposed method provides a better fault injection resilience indicator compared to fault-tolerance training on average by 5%. Also, the proposed method provides a better resilience to evasion adversarial attack compared to adversarial training on average by 4.8%. It has also demonstrated an average improvement by 4.8% in task change resilience compared to conventional fine-tuning of tuner blocks.

The results of the conventional MAML and the Proposed Resilient-aware MAML are compared in terms of the impact on resilience to disturbances. The advantage of the proposed method is confirmed. In addition, meta-learning with an evolutionary strategy provides on average higher values of the resilience indicator, although it requires more iterations.

The practical significance of the achieved results lies in the formation of a new methodological basis for the development of algorithms for optimizing the resilience of AIS, which is important for safety-critical applications. Moreover, the method has a fairly unified structure and can be applied to a wide range of AIS model architectures and tasks, which brings it closer to the concept of providing resiliency as a service.

The limitations of the research are related to testing this approach only on the ResNet-18 convolutional network with blocks of tuners and meta-tuners based on the Conv-Adapter architecture. Nevertheless, the paper shows the fundamental possibility of increasing the resilience of the original model by using tuners, meta-tuners, and meta-learning with a perturbation generator.

Future research should focus on the architecture of the tuner blocks and the application of the proposed approach to other machine learning tasks, such as regression, reinforcement learning, and generative models.

ACKNOWLEDGMENTS

The research was concluded in the Intellectual Systems Laboratory of Computer Science Department at Sumy State University with the financial support of the Ministry of Education and Science of Ukraine in the framework of state budget scientific and research work of DR No. 0122U000782 “Information technology for providing resilience of artificial intelligence systems to protect cyber-physical systems”.

REFERENCES

1. Moskalenko V., Kharchenko V., Moskalenko A., and Kuzikov B. Resilience and Resilient Systems of Artificial Intelligence: Taxonomy, models and methods, *Algorithms*, Vol. 16, No. 3, pp. 1–44, 2023. DOI:10.3390/a16030165.
2. Chakraborty A., Alam M., Dey V., Chattopadhyay A. and D. Mukhopadhyay, *A survey on adversarial attacks and defences*, *CAAI Transactions on Intelligence Technology*, 2021.Vol. 6, No. 1, pp. 25–45. DOI:10.1049/cit2.12028.
3. Hoang L.-H., Hanif M. A. and Shafique M. Tre-map: Towards reducing the overheads of fault-aware retraining of deep neural networks by merging fault maps, *Proceedings of the 2021 24th Euromicro Conference on Digital System Design (DSD)*. Palermo, Italy, 1–3 September 2021, 8 p. DOI:10.1109/dsd53832.2021.00072.
4. Lu J., Liu A., Dong F., Gu F., Gama J. and Zhang G., Learning under Concept Drift: A Review, *IEEE Transactions on Knowledge and Data Engineering*, 2019, Vol. 31, No. 12, pp. 2346–2363, DOI: 10.1109/TKDE.2018.2876857.
5. Dymond J., Graceful degradation and related fields, ePrints Soton, https://eprints.soton.ac.uk/455349/ (accessed May 18, 2023).
6. Eigner O., Xu K., Liu S., Chen Pin-Yu, Weng Tsui-Wei, Gan Ch. and Wang M., Towards resilient artificial intelligence: Survey and research issues, *2021 IEEE International Conference on Cyber Security and Resilience (CSR)*, 2021. Rhodes, Greece, 26–28 July 2021. pp. 536–542. DOI:10.1109/csr51186.2021.9527986.
7. Wang R., Xu K. , Liu S., Chen Pin-Yu et al. On Fast Adversarial Robustness Adaptation in Model-Agnostic Meta-Learning, ArXiv, Vol. abs/2102.10454, 2021, pp. 1–16. DOI: 10.48550/arXiv.2102.10454.
8. Son X., Yang Y., Choromanski K., Caluwaerts K., Gao W., Finn C. and Tan J. Rapidly adaptable legged robots via evolutionary meta-learning, *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Las Vegas, NV, USA, 24 October 2020 – 24 January 2021, pp. 1–11. DOI:10.1109/iros45743.2020.9341571.
9. Ding N., Qin Y., Yang G. et al. Parameter-efficient fine-tuning of large-scale pre-trained language models, *Nature Machine Intelligence*, 2023, Vol. 5, No. 3, pp. 220–235. DOI:10.1038/s42256-023-00626-4.
10. Fraccascia L., Giannoccaro I., and Albino V. Resilience of Complex Systems: State of the art and directions for future

DOI 10.15588/1607-3274-2023-2-9

research, *Complexity*, 2018, Vol. 2018, pp. 1–44. DOI:10.1155/2018/3421529.
11. Drozd O., Kharchenko V., Rucinski A., Kochanski T., Garbos R. and Maevsky D. Development of models in resilient computing, *2019 10th International Conference on Dependable Systems, Services and Technologies (DESSERT)*, 2019, Leeds, United Kingdom, 5–7 June 2019, pp. 1–6. DOI:10.1109/dessert.2019.8770035.
12. Inouye B. D., Brosi B. J., Le Sage E. H., and M. T. Lerdau, Trade-offs among resilience, robustness, stability, and performance and how we might study them, *Integrative and Comparative Biology*, 2021, Vol. 61, No. 6, pp. 2180–2189. DOI:10.1093/icb/icab178.
13. Santos S. G. T. d. C., Gonçalves Júnior P. M. , Silva G. D. d. S. and de Barros R. S. M. Speeding Up Recovery from Concept Drifts, *Machine Learning and Knowledge Discovery in Databases*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 179–194. DOI: 10.1007/978-3-662-44845-8_12.
14. Xie C., Wang J., Zhang Z., Ren Z. and Yuille A. Mitigating Adversarial Effects Through Randomization, *Proceedings of the International Conference on Learning Representations*, Toulon. France, 24–26 April 2017, pp. 1–16. DOI: 10.48550/arXiv.1711.01991.
15. Kwon H., and Lee J. Diversity Adversarial Training against Adversarial Attack on Deep Neural Networks, *Symmetry*, 2021, Vol. 13, No. 3, pp. 1–14, DOI: 10.3390/sym13030428.
16. Abusnaina A., Wu Y., Arora S. et al. Adversarial example detection using latent neighborhood graph, *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada, 10–17 October 2021, pp. 7687–7696. DOI:10.1109/iccv48922.2021.00759.
17. Li W., Ning X., Ge G., Chen X., Wang Y. and Yang H. FTT-NAS: Discovering Fault-Tolerant Neural Architecture, *2020 25th Asia South Pacific Des. Automat. Conf. (ASP-DAC)*. Beijing, China, 13–16 Jan. 2020. IEEE, 2020, pp. 2011–2016. DOI: 10.1109/asp-dac47756.2020.9045324.
18. Xu H., Chen Z., Wu W., Jin Z., Kuo S.-Y. and Lyu M. NV-DNN: Towards Fault-Tolerant DNN Systems with N-Version Programming, *2019 49th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw. Workshops (DSN-W)*. Portland, OR, USA, 24–27 Jun. 2019. IEEE, 2019, pp. 44–47. DOI: 10.1109/dsn-w.2019.00016.
19. Javaheripi M. and Koushanfar F. HASHTAG: Hash Signatures for Online Detection of Fault-Injection Attacks on Deep Neural Networks, *2021 IEEE/ACM Int. Conf. Comput. Aided Des. (ICCAD)*. Munich, Germany, 1–4 Nov. 2021. IEEE, 2021, pp. 1–9. DOI: 10.1109/iccad51958.2021.9643556.
20. Volpi R., Namkoong H., Sener O.et al. Generalizing to unseen domains via adversarial data augmentation, *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Montréal, QC, Canada, 2–8 December 2018, pp. 1–11. DOI: 10.5555/3327345.3327439.
21. Kulinchenko H. V., Drozdenko O. O., Leontiev P. V. and Hrek V. M. Pressure Regulator for Low Temperature Separation Process, *2021 IEEE 12th Int. Conf. Electron. Inf. Technol. (ELIT)*. Lviv, Ukraine, 19–21 May 2021. IEEE, 2021, pp. 315–319. DOI: 10.1109/elit53502.2021.9501143.
22. An W., Wang H. , Sun Q., Xu J., Dai Q. and Zhang L. A PID controller approach for stochastic optimization of Deep Networks, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, Salt Lake City, UT, 18–23 June 2018, pp. 8522–8531. DOI:10.1109/cvpr.2018.00889.
23. Tian Y., Zhao X. and Huangm W. Meta-learning approaches for learning-to-learn in deep learning: A survey, Neurocomputing, 2022, Vol. 494, pp. 203–223. DOI: 10.1016/j.neucom.2022.04.078.
24. Yang X. and Xu J., Few-shot Classification with First-order Task Agnostic Meta-learning, *2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA)*. Changchun, China, 20–22 May 2022, pp. 2017–2020. DOI:10.1109/cvidliccea56201.2022.9824307.
25. Song X., Yang Y., Choromanski K. et al. Rapidly Adaptable Legged Robots via Evolutionary Meta-Learning, *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Las Vegas, NV, USA, 24 October 2020 – 24 January 2021, pp. 3769–3776. DOI:10.1109/iros45743.2020.9341571.
26. Wheaton M. and Madni Azad M. Resiliency and Affordability Attributes in a System Tradespace, *AIAA SPACE 2015 Conference and Exposition*, Pasadena, California. Reston, Virginia, 31 Aug-2 Sep 2015. DOI:10.2514/6.2015-4434.
27. Bansal T., Alzubi S., Wang T., Lee Jay-Yoon and McCallum A., Meta-Adapters: Parameter Efficient Few-shot Fine-tuning through Meta-Learning, *First Conference on Automated Machine Learning*, 2022, 18 p. Available at: https://openreview.net/pdf?id=bQt8dWKsfso.
28. Jiang Z., Jiang Z., Mao Ch. et al. Rethinking Efficient Tuning Methods from a Unified Perspective (Version 1), *arXiv*, 2023. DOI:10.48550/ARXIV.2303.00690.
29. Chen H., Tao R., Zhang H. et al. Conv-Adapter: Exploring Parameter Efficient Transfer Learning for ConvNets (Version 3), *arXiv*, 2022. DOI:10.48550/ARXIV.2208.07463.
30. Wu N., Hou H., Jia X., Chang X. and Li H. Low-Resource Neural Machine Translation Based on Improved Reptile Meta-learning Method, *Communications in Computer and Information Science*. Singapore, 2021. pp. 39–50. DOI: 10.1007/978-981-16-7512-6_4.
31. Park S. and So J., On the Effectiveness of Adversarial Training in Defending against Adversarial Example Attacks for Image Classification, *Applied Sciences*, 2020, Vol. 10, No. 22, pp. 1–16. DOI: 10.3390/app10228079.
32. Kotyan Sh., and Vargas D. Vasconcellos, Adversarial robustness assessment: Why in evaluation both L0 and L∞ attacks are necessary, PLOS ONE, 2022, Vol. 17, No. 4, pp. e0265723, DOI: 10.1371/journal.pone.0265723.
33. Li G., Pattabiraman K.and DeBardeleben N. TensorFI: A Configurable Fault Injector for TensorFlow Applications, *2018 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*. Memphis, TN, 15–18 October 2018, pp. 1–8. DOI: 10.1109/issrew.2018.00024.
34. Foldy-Porto T., Venkatesha Y. and Panda P. Activation Density Driven Efficient Pruning in Training, *2020 25th International Conference on Pattern Recognition (ICPR)*, Milan, Italy, 10–15 January 2021, pp. 8929–8936. DOI: 10.1109/icpr48806.2021.9413182.

Received 05.05.2023.
Accepted 31.05.2023.

DOI 10.15588/1607-3274-2023-2-9

УДК 004.891.032.26:629.7.01.066

НЕЗАЛЕЖНИЙ ВІД МОДЕЛІ АЛГОРИТМ МЕТА-НАВЧАННЯ ДЛЯ ОПТИМІЗАЦІЇ РЕЗІЛЬЄНТНОСТІ СИСТЕМИ ШТУЧНОГО ІНТЕЛЕКТУ

Москаленко В. В. – канд. техн. наук, доцент, доцент кафедри комп'ютерних наук, Сумський державний університет, Суми, Україна.

АНОТАЦІЯ

Актуальність. Задача оптимізації резільєнтності систем штучного інтелекту до деструктивних збурень досі не була повністю вирішена і є досить актуальною для критичних до безпеки застосувань. Задача оптимізації резільєнтності системи штучного інтелекту до збурюючих впливів є високорівневою по відношенню до оптимізації ефективності, що обумовлює перспективність використання ідей і методів мета-навчання для її вирішення. Тому об'єктом дослідження є процес мета-навчання для оптимізації резільєнтності системи штучного інтелекту до деструктивних збурень. Предметом дослідження є архітектурні надстройки та метод мета-навчання, що забезпечують оптимізацію резільєнтності до протиборчих атак, інжекції несправностей і зміни задач.

Мета дослідження – розроблення ефективного методу мета-навчання для оптимізації резільєнтності системи штучного інтелекту до деструктивних збурень.

Методи дослідження. Оптимізація резільєнтності реалізується шляхом поєднання ідей і методів протиборчого навчання, навчання з ін'єкцією несправностей, незалежного від моделі мета-навчання, навчання за обмеженою кількістю зразків, методів градієнтної оптимізації та ймовірнісних стратегій апроксимації градієнту. При цьому вибір архітектурних надстройок базується на ефективному щодо параметрів трансфері знань для для економії ресурсів та уникнення проблеми катастрофічного забування.

Результати. Розроблено незалежний від моделі метод мета-навчання для оптимізації резільєнтності систем штучного інтелекту на основі градієнтних мета-оновлень, або мета-оновлень за еволюційною стратегією. При цьому метод передбачає використання тюнерів і мета-тюнерів, що здійснюють паралельну корекцію будівельних модулів (блоків) глибокої нейромережі. На прикладі задачі класифікації зображень експериментально протестовано здатність запропонованого підходу підвищувати ефективність поглинання збурень та підвищувати інтегральний показник резільєнтності системи штучного інтелекту. Експерименти проводились на моделі з архітектурою ResNet-18, з надстройкою у вигляді тюнерів і мета-тюнерів з архітектурою Conv-Adapter. При цьому CIFAR-10 використовується як базовий набір, на якому була навчена модель, а CIFAR-100 використовується як набір для формування вибірок, на яких здійснюють адаптацію за обмеженою кількістю зразків. Порівнюється показники резільєнтності системи штучного інтелекту після попереднього навчання тюнерів і мета-тюнерів за алгоритмом протиборчого навчання, алгоритмом навчання з ін'єкцією несправностей, традиційним алгоритмом незалежного від моделі мета-навчання та за запропонованим метод мета-навчання для оптимізації резільєнтності. Також порівнюються за інтегральним показником резільєнтності алгоритм мета-навчання з мета-градієнтним оновленням та мета-оновленням на основі еволюційної стратегії.

Висновки. Есперементально підтверджено, що запропонований метод забезпечує кращий показник резільєнтності до ін'єкції випадкових інверсій біт порівняно з навчанням з ін'єкцією несправностей в середньому на 5%. Також запропонований метод забезпечує кращий показник резільєнтності до L_∞ протиборчих атак ухилення порівняно з протиборчим навчання всередньому на 4.8%. Так само продемонстровано підвищення всередньому на 4.8% резільєнтності до зміни задач порівняно зі звичайною точною настройкою тюнерів. При цьому мета-навчання з еволюційною стратегією забезпечує всередньому більші значення показника резільєнтності, однак попереднє мета-навчання потребує більше ітерацій.

КЛЮЧОВІ СЛОВА: мета-навчання, еволюційна стратегія, ефективний стосовно параметрів трансфер знань, робастність, резільєнтність, протиборчі атаки, інжекція несправностей, навчання з декількох зразків.

ЛІТЕРАТУРА

1. Resilience and Resilient Systems of Artificial Intelligence: Taxonomy, Models and Methods / [V. Moskalenko, V. Kharchenko A. Moskalenko, B. Kuzikov] // Algorithms. – 2023. – Vol. 16, No. 3. – P. 1–44. DOI: 10.3390/a16030165.
2. A survey on adversarial attacks and defences / [A. Chakraborty, M. Alam, V. Dey et al.] // CAAI Transactions on Intelligence Technology. – 2021. – Vol. 6, No. 1. – P. 25–45. DOI: 10.1049/cit2.12028.
3. Hoang L.-H. TRe-Map: Towards Reducing the Overheads of Fault-Aware Retraining of Deep Neural Networks by Merging Fault Maps / Le-Ha Hoang, M. Abdullah Hanif, M. Shafique // Digital System Design (DSD) : 24th Euromicro Conference, Palermo, Italy, 1–3 September 2021 : proceedings. – 8 p. DOI: 10.1109/dsd53832.2021.00072.
4. Learning under Concept Drift: A Review / [J. Lu, A. Liu, F. Dong et al.] // IEEE Transactions on Knowledge and Data Engineering. – 2018. – Vol. 31, No. 12. – P. 2346–2363. DOI: 10.1109/tkde.2018.2876857.
5. Dymond J. Graceful Degradation and Related Fields-ePrints Soton. Welcome to ePrints Soton-ePrints Soton [Electronic resource]. – Access mode: https://eprints.soton.ac.uk/455349/.
6. Towards Resilient Artificial Intelligence: Survey and Research Issues / [O. Eigner, S. Eresheim, P. Kieseberg et al.] // Cyber Security and Resilience (CSR) : IEEE International Conference, Rhodes, Greece, 26–28 July 2021 : proceedings. – P. 536–542. DOI: 10.1109/csr51186.2021.9527986.
7. On Fast Adversarial Robustness Adaptation in Model-Agnostic Meta-Learning / [R. Wang, K. Xu, S. Liu, Pin-Yu Chen et al.] // ArXiv. – Vol. abs/2102.10454. – 2021. – P. 1–16. DOI: 10.48550/arXiv.2102.10454.
8. Rapidly Adaptable Legged Robots via Evolutionary Meta-Learning / [X. Song, Y. Yang, K. Choromanski et al.] // Intelligent Robots and Systems (IROS) : IEEE/RSJ International Conference, Las Vegas, NV, USA, 24 October 2020 – 24 January 2021 : proceedings. – P. 1–11. DOI: 10.1109/iros45743.2020.9341571.
9. Parameter-efficient fine-tuning of large-scale pre-trained language models / [N. Ding, Y. Qin, G. Yang et al.] // Nature Machine Intelligence. – 2023. – Vol. 5. – P. 220–235. DOI: 10.1038/s42256-023-00626-4.

DOI 10.15588/1607-3274-2023-2-9

10. Fraccascia L. Resilience of Complex Systems: State of the Art and Directions for Future Research / L. Fraccascia, I. Giannoccaro, V. Albino // Complexity. – 2018. – Vol. 2018. – P. 1–44 DOI: 10.1155/2018/3421529.
11. Development of Models in Resilient Computing / [O. Drozd, V. Kharchenko, A. Rucinski et al.] // Dependable Systems, Services and Technologies (DESSERT) : 10th International Conference, Leeds, United Kingdom, 5–7 June 2019 : proceedings. – P. 1–6. DOI: 10.1109/dessert.2019.8770035.
12. Trade-offs Among Resilience, Robustness, Stability, and Performance and How We Might Study Them / [Brian D. Inouye, Berry J. Brosi, Emily H. Le Sage, M. T. Lerdau] // Integrative and Comparative Biology. – 2021. – Vol. 61, No. 6. – P. 2180–2189. DOI:10.1093/icb/icab178.
13. Speeding Up Recovery from Concept Drifts / [S. G.Teixeira de Carvalho Santos, P. M. Gonçalves Júnior, G. D. dos Santos Silva et al.] // Machine Learning and Knowledge Discovery in Databases. – Berlin, Heidelberg, 2014 : proceedings. – P. 179–194. DOI: 10.1007/978-3-662-44845-8_12.
14. Mitigating Adversarial Effects Through Randomization / [C. Xie, J. Wang, Z. Zhang et al.] // Learning Representations : International Conference, Toulon, France, 24–26 April 2017 : proceedings. – Openreview.net, 2017. – P. 1–16. DOI: 10.48550/arXiv.1711.01991.
15. Kwon H. Diversity Adversarial Training against Adversarial Attack on Deep Neural Networks / H. Kwon, J. Lee // Symmetry. – 2021. – Vol. 13, no. 3. – P. 1–14. DOI: 10.3390/sym13030428.
16. Adversarial Example Detection Using Latent Neighborhood Graph / [A. Abusnaina, Y. Wu, S. Arora et al.] // Computer Vision (ICCV) : IEEE/CVF International Conference, Montreal, QC, Canada, 10–17 October 2021 : proceedings. – P. 7687–7696. DOI: 10.1109/iccv48922.2021.00759.
17. FTT-NAS: Discovering Fault-Tolerant Neural Architecture / [W. Li, X. Ning, X. Ning et al.] // 2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC), Beijing, China, 13–16 January 2020 : proceedings. – P. 2011–2016. DOI: 10.1109/asp-dac47756.2020.9045324.
18. NV-DNN: Towards Fault-Tolerant DNN Systems with N-Version Programming / [H. Xu, Zh. Chen, W. Wu et al.] // Dependable Systems and Networks Workshops (DSN-W) : 49th Annual IEEE/IFIP International Conference, Portland, OR, USA, 24–27 June 2019 : proceedings. – P. 44–47. DOI: 10.1109/dsn-w.2019.00016.
19. Javaheripi M. HASHTAG: Hash Signatures for Online Detection of Fault-Injection Attacks on Deep Neural Networks / M. Javaheripi, F. Koushanfar // Computer Aided Design (ICCAD) : IEEE/ACM International Conference, Munich, Germany, 1–4 November 2021 : proceedings. – P. 1–9. DOI: 10.1109/iccad51958.2021.9643556.
20. Generalizing to unseen domains via adversarial data augmentation / [R. Volpi, H. Namkoong, O. Sener et al.] // Neural Information Processing Systems : 32nd International Conference, Montréal, QC, Canada, 2–8 Dec. 2018 : proceedings. – P. 1–11. DOI: 10.5555/3327345.3327439.
21. Pressure Regulator for Low Temperature Separation Process / [H. V. Kulinchenko, O. O. Drozdenko, P. V. Leontiev et al.] // Electronics and Information Technologies (ELIT) : IEEE 12th International Conference, Lviv, Ukraine, 19–21 May 2021 : proceedings. – IEEE, 2021. – P. 315–319. DOI: 10.1109/elit53502.2021.9501143.
22. A PID Controller Approach for Stochastic Optimization of Deep Networks / [W. An, H. Wang, Q. Sun et al.] // Computer Vision and Pattern Recognition (CVPR) : IEEE/CVF Conference, Salt Lake City, UT, 18–23 June 2018 : proceedings. – IEEE, 2018. – P. 8522–8531. DOI: 10.1109/cvpr.2018.00889.
23. Tian Y. Meta-learning approaches for learning-to-learn in deep learning: A survey / Y. Tian, X. Zhao, W. Huang // Neurocomputing. – 2022. – Vol. 494. – P. 203–223. DOI: 10.1016/j.neucom.2022.04.078.
24. Yang X. Few-shot Classification with First-order Task Agnostic Meta-learning / X. Yang, J. Xu // Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA) : 3rd International Conference, Changchun, China, 20–22 May 2022 : proceedings. – Los Alamitos, IEEE, 2022. – P. 2017–2020 DOI:10.1109/cvidliccea56201.2022.9824307.
25. Rapidly Adaptable Legged Robots via Evolutionary Meta-Learning / [X. Song, Y. Yang, K. Choromanski et al.] // Intelligent Robots and Systems (IROS) : IEEE/RSJ International Conference, Las Vegas, NV, USA, 24 October 2020 – 24 January 2021 : proceedings. – P. 3769–3776. DOI:10.1109/iros45743.2020.9341571.
26. Wheaton M. Resiliency and Affordability Attributes in a System Tradespace / M. Wheaton, Azad M. Madni // AIAA SPACE 2015 Conference and Exposition, Pasadena, California, 31 Aug-2 Sep., 2015 : proceedings. DOI:10.2514/6.2015-4434.
27. Meta-Adapters: Parameter Efficient Few-shot Fine-tuning through Meta-Learning / [T. Bansal, S. Alzubi, T. Wang et al.] // Automated Machine Learning : First Conference : proceedings. – 2022. – P. 18. – Access mode : https://openreview.net/pdf?id=bQt8dWKsfso.
28. Rethinking Efficient Tuning Methods from a Unified Perspective (Version 1) / [Z. Jiang, Z. Jiang, Ch. Mao et al.] // arXiv. – 2023. DOI:10.48550/ARXIV.2303.00690.
29. Conv-Adapter: Exploring Parameter Efficient Transfer Learning for ConvNets (Version 3) / [H. Chen, R. Tao, H. Zhang et al.] // arXiv. – 2022. – DOI:10.48550/ARXIV.2208.07463.
30. Low-Resource Neural Machine Translation Based on Improved Reptile Meta-learning Method / [N. Wu, H. Hou, X. Jia et al.] // Communications in Computer and Information Science. – Singapore, 2021. – P. 39–50. DOI: 10.1007/978-981-16-7512-6_4.
31. Park S. On the Effectiveness of Adversarial Training in Defending against Adversarial Example Attacks for Image Classification / S. Park, J. So // Applied Sciences. – 2020. – Vol. 10, no. 22. – P. 1–16. – DOI: 10.3390/app10228079.
32. Kotyan Sh. Adversarial robustness assessment: Why in evaluation both L0 and L∞ attacks are necessary / Sh. Kotyan, D. Vasconcellos Vargas // PLOS ONE. – 2022. – Vol. 17, no. 4. – P. e0265723. DOI: 10.1371/journal.pone.0265723.
33. Li G. TensorFI: A Configurable Fault Injector for TensorFlow Applications / G. Li, K. Pattabiraman, N. DeBardeleben // Software Reliability Engineering Workshops (ISSREW) : IEEE International Symposium, Memphis, TN, 15–18 October 2018 : proceedings. – Los Alamitos, IEEE, 2018. – P. 1–8. DOI: 10.1109/issrew.2018.00024.
34. Foldy-Porto T. Activation Density Driven Efficient Pruning in Training / T. Foldy-Porto, Y. Venkatesha, P. Panda // Pattern Recognition (ICPR): 25th International Conference , Milan, Italy, 10–15 January 2021 : proceedings – P. 8929–8936. DOI: 10.1109/icpr48806.2021.9413182.

DOI 10.15588/1607-3274-2023-2-9