https://doi.org/10.31891/2307-5732-2025-351-44
УДК 004.75:048

ПЕРЕТЯГА МАКСИМ
Харківський національний університет радіоелектроніки
https://orcid.org/0000-0002-9675-1305
e-mail: maksym.peretiaha@nure.ua
РЕВЕНЧУК ІЛОНА
Харківський національний університет радіоелектроніки
https://orcid.org/0000-0002-5188-9538
e-mail: ilona.revenchuk@nure.ua

ПОРІВНЯЛЬНИЙ АНАЛІЗ МЕТОДІВ ДЛЯ ВИЯВЛЕННЯ АНОМАЛІЙ У МІКРОСЕРВІСАХ ІЗ ЗАСТОСУВАННЯМ НЕЙРОННИХ МЕРЕЖ

Предметом дослідження є застосування нейронних мереж для виявлення аномалій у мікросервісних архітектурах. Мета роботи полягає у порівнянні ефективності різних нейронних підходів до виявлення аномалій, визначенні їх переваг та недоліків, а також обґрунтуванні вибору оптимального методу. Для досягнення мети визначено наступні завдання: огляд сучасних методів виявлення аномалій у мікросервісах; аналіз підходів, заснованих на нейронних мережах, зокрема автоенкодерах та рекурентних мережах; порівняння методів за критеріями ефективності, обчислювальної складності та стійкості до шуму; аналіз можливостей комбінування методів для підвищення точності виявлення аномалій; формулювання рекомендацій щодо вибору оптимального підходу та перспектив подальших досліджень. Розглянуто метод використання глибоких нейронних мереж для аналізу поведінкових аномалій у мікросервісних системах. Результати дослідження показали, що глибокі нейронні мережі демонструють високу точність у виявленні аномалій, при цьому різні архітектури мають відмінності в швидкодії та стійкості до помилкових спрацьовувань. Висновки свідчать про ефективність застосування нейронних мереж у задачах моніторингу мікросервісів, а також про необхідність подальшого дослідження комбінованих підходів для підвищення точності та продуктивності систем виявлення аномалій.

Ключові слова: нейронні мережі, виявлення аномалій, мікросервіс, автоенкодер.

PERETIAHA MAKSYM, REVENCHUK ILONA
Kharkiv National University of Radio Electronics, Ukraine

APPLICATION OF NEURAL NETWORKS TO DETECT ANOMALIES IN MICROSERVICES: COMPARATIVE ANALYSIS OF METHODS.

This study focuses on the application of neural networks for anomaly detection in microservice architectures. The increasing complexity of modern microservice systems makes traditional anomaly detection methods less effective, as they struggle to adapt to dynamic environments and large volumes of data. Neural networks offer a promising solution due to their ability to model complex patterns and detect subtle deviations that may indicate system failures or security threats.

The aim of this work is to compare the effectiveness of different neural network-based approaches to anomaly detection, evaluate their strengths and weaknesses, and determine the most suitable method for practical implementation. To achieve this, several key tasks were outlined: an in-depth review of contemporary anomaly detection techniques applied to microservices; an analysis of neural network-based models, with a particular focus on autoencoders and recurrent neural networks (RNNs); a comparative assessment of these methods in terms of accuracy, computational efficiency, and robustness to noise; an exploration of hybrid models that combine different techniques to improve detection performance; and the formulation of recommendations for selecting the optimal approach based on specific application scenarios.

This research examines the use of deep neural networks for detecting behavioral anomalies in microservice environments. Various architectures are tested and compared to identify trade-offs between detection accuracy, processing speed, and resistance to false positives. The findings indicate that deep neural networks, particularly autoencoders and recurrent models, achieve high precision in identifying anomalies, but their efficiency varies depending on the architecture and the nature of the dataset. The study highlights that while autoencoders effectively capture deviations in data distribution, recurrent networks excel in detecting temporal anomalies in sequential microservice interactions.

The conclusions confirm that neural networks provide a powerful tool for anomaly detection in microservice monitoring, significantly outperforming conventional statistical approaches. However, challenges remain, including the need for adaptive models capable of learning in real time and optimizing computational resources. Further research should focus on hybrid approaches that integrate multiple neural architectures to enhance detection accuracy and scalability. The development of self-learning and auto-tuning models could also improve adaptability to evolving microservice behaviors, ensuring robust and efficient anomaly detection in complex distributed systems.

Keywords: neural networks, anomaly detection, microservice, autoencoder.

Стаття надійшла до редакції / Received 03.03.2025
Прийнята до друку / Accepted 18.04.2025

Постановка проблеми

Мікросервісна архітектура широко використовується для створення гнучких і масштабованих програмних систем. Однак велика кількість взаємодіючих сервісів ускладнює моніторинг та діагностику аномалій, що можуть призводити до збоїв, зниження продуктивності або порушення безпеки. Традиційні методи виявлення аномалій, такі як статистичні підходи та правила експертних систем, часто виявляються недостатньо ефективними в умовах високої динамічності мікросервісного середовища. Вони погано адаптуються до нових патернів поведінки, мають високу обчислювальну

складність та схильні до високого рівня помилкових спрацьовувань. Застосування нейронних мереж для виявлення аномалій здатні автоматично навчатися складним залежностям у даних, виявляти приховані закономірності та адаптуватися до змін у системі. Проте різні архітектури нейронних мереж мають різні характеристики, що впливає на їхню ефективність у конкретних сценаріях моніторингу мікросервісів. Таким чином, постає проблема вибору оптимального методу на основі нейронних мереж для виявлення аномалій у мікросервісних системах. Необхідний детальний аналіз існуючих підходів, їхніх переваг та недоліків, а також визначення критеріїв, за якими можна оцінювати їхню ефективність у реальних умовах експлуатації.

Аналіз досліджень та публікацій

Проблема виявлення аномалій у розподілених системах, зокрема у мікросервісній архітектурі, є предметом численних досліджень. Традиційні підходи включають статистичні методи, машинне навчання та нейронні мережі.

Статистичні методи, такі як Z-score, контрольні карти Шухарта та методи Байєсового висновку, широко застосовуються для моніторингу змін у поведінці систем [1]. Проте їх ефективність значно знижується у складних та динамічних середовищах через необхідність попереднього визначення розподілу даних та високий рівень хибнопозитивних спрацьовувань [2].

Методи машинного навчання, такі як метод опорних векторів (SVM), алгоритми випадкових лісів та методи кластеризації (наприклад, k-means), дозволяють виявляти аномалії на основі аналізу характеристик поведінки мікросервісів [3]. Однак ці підходи вимагають великого обсягу апріорних даних для навчання, що обмежує їхню адаптивність у динамічних системах [4].

У цьому контексті нейронні мережі пропонують більш гнучкий підхід, оскільки здатні навчатися складним закономірностям у даних та адаптуватися до змін у поведінці мікросервісів [5].

Нейронні мережі демонструють високу ефективність у завданнях виявлення аномалій завдяки здатності моделювати нелінійні залежності та виявляти приховані патерни в потоках даних. Основними підходами є автоенкодери, рекурентні нейронні мережі (RNN), згорткові нейронні мережі (CNN) та гібридні моделі [6].

Автоенкодери широко використовуються для аномального аналізу, оскільки навчаються реконструювати вхідні дані та можуть визначати аномалії на основі відхилень між вхідним та відновленим сигналами [7]. У мікросервісних системах автоенкодери ефективні при виявленні нестандартних патернів навантаження, проте їх продуктивність залежить від розмірності простору ознак та якості вибору гіперпараметрів [8].

Рекурентні нейронні мережі (RNN), зокрема довготривало-короткочасна пам’ять (LSTM), застосовуються для аналізу часових рядів, що дозволяє ефективно виявляти аномалії у поведінці мікросервісів на основі історичних даних [9]. Порівняно з автоенкодерами, LSTM забезпечують кращу здатність прогнозувати майбутні стани системи, однак мають високу обчислювальну складність, що може бути критичним для реального часу [10].

Згорткові нейронні мережі (CNN**)**, хоча зазвичай застосовуються у сфері комп’ютерного зору, також використовуються для виявлення аномалій у системах з багатовимірними вхідними даними, наприклад у мережевому моніторингу мікросервісів [11]. Їхня ключова перевага – здатність виділяти просторові закономірності у вхідних даних, однак вони менш ефективні для аналізу часових рядів порівняно з RNN та LSTM [12].

Згідно з дослідженням [13], автоенкодери демонструють високу чутливість до нових типів аномалій, але схильні до переобучення, якщо дані містять значну кількість шуму. LSTM-моделі ефективні для виявлення довготривалих змін у поведінці системи, проте їхнє навчання потребує значних обчислювальних ресурсів. CNN мають швидку продуктивність, однак поступаються RNN у точності виявлення аномалій у часових даних.

Дослідження [14] показує, що комбіновані підходи, такі як гібридні автоенкодери з рекурентними шарами або ансамблеві методи, можуть забезпечити вищу точність та стійкість до помилкових спрацьовувань. Зокрема, використання ансамблів автоенкодерів та рекурентних мереж дозволяє комбінувати переваги обох підходів, підвищуючи ефективність виявлення аномалій у реальному часі [15].

Огляд літератури показує, що нейронні мережі є перспективним інструментом для виявлення аномалій у мікросервісних системах. Автоенкодери добре підходять для виявлення нетипових патернів у великих потоках даних, рекурентні мережі ефективні для аналізу часових рядів, а згорткові мережі можуть застосовуватися для обробки складних багатовимірних вхідних даних.

Проте кожен метод має свої обмеження, і універсального підходу не існує. Це вказує на необхідність подальших досліджень у напрямку гібридних моделей та розробки адаптивних алгоритмів, що можуть ефективно працювати у різних сценаріях експлуатації мікросервісних систем.

Метою роботи є дослідження і порівняння нейронних підходів до виявлення аномалій, їх переваг та недоліків, а також обґрунтування вибору оптимального.

Виклад основного матеріалу

1. Загальні підходи до виявлення аномалій у мікросервісних системах

Виявлення аномалій у мікросервісних архітектурах є важливим завданням для забезпечення стабільності та продуктивності програмних систем. Основні труднощі полягають у високій динамічності навантаження, варіативності запитів та великій кількості взаємодій між сервісами. У зв'язку з цим використовуються різні методи, які можна умовно розділити на три основні групи:

- статистичні методи – базуються на оцінці відхилень у поведінці системи, використовуючи такі підходи, як Z-score, методи Байєсового висновку та контрольні карти Шухарта [1]. Основні недоліки цих методів – їхня чутливість до змін у розподілі даних та необхідність попереднього визначення нормальної поведінки;
- методи машинного навчання – передбачають використання алгоритмів кластеризації (наприклад, k-means), дерев рішень, випадкових лісів та методів опорних векторів (SVM) для ідентифікації аномалій [2]. Вони можуть бути як контрольованими, так і неконтрольованими, однак вимагають великого обсягу навчальних даних та значних обчислювальних ресурсів;
- методи на основі нейронних мереж – дозволяють будувати нелінійні моделі, здатні самостійно виявляти приховані закономірності в даних. Найбільш перспективними у контексті мікросервісів є автоенкодери, рекурентні нейронні мережі (RNN), згорткові нейронні мережі (CNN) та їхні комбінації [3].

2. Використання нейронних мереж для виявлення аномалій

Методи, засновані на нейронних мережах, забезпечують кращу гнучкість у порівнянні зі статистичними підходами та традиційним машинним навчанням. Вони не потребують ручного визначення граничних значень, можуть працювати з великими наборами даних та адаптуватися до змін у поведінці системи. У цьому розділі розглянуто основні архітектури нейронних мереж, що використовуються для виявлення аномалій у мікросервісах.

2.1 Автоенкодери

Автоенкодери є одним із найпоширеніших підходів до виявлення аномалій, оскільки дозволяють навчати модель реконструювати нормальні зразки даних, а відхилення у відновленні сигналу можуть використовуватися для визначення аномалій [4]. На рисунку 1 наведено схематичний принцип роботи автоенкодера:

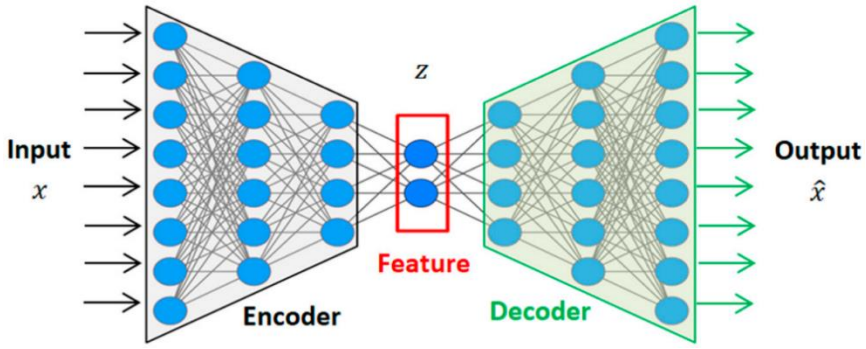

Рис.1. Схематичний принцип роботи автоенкодера

Основна ідея автоенкодерів полягає у створенні вузького латентного представлення вхідних даних, що дозволяє мережі виявляти приховані закономірності. Архітектура автоенкодера складається з двох частин:

- енкодер – стискає вхідні дані до меншого за розміром латентного простору;
- декодер – відновлює вхідні дані з латентного представлення.

Навчання автоенкодера відбувається шляхом мінімізації функції втрат, яка визначає різницю між вхідними даними та їх реконструкцією. Для цього зазвичай використовують середньоквадратичну похибку (MSE):

$$L = \frac{1}{n}\sum_{i-1}^{n}(x_i - x\hat{}_i)^2 \quad (1),$$

де x_i – вхідні дані, $x\hat{}_i$ – відновлені дані, nnn – розмір вибірки.

Якщо автоенкодер навчається на нормальних даних, то він добре відтворює їх, тоді як аномальні зразки дають значно більшу похибку реконструкції, що дозволяє використовувати цю похибку як критерій виявлення аномалій.

Типова архітектура автоенкодера складається з кількох шарів нейронів у енкодері та декодері. Основні компоненти:

- вхідний шар: приймає вхідні дані у вигляді векторів.
- приховані шари енкодера: виконують поступове зменшення розмірності даних.
- латентний простір: містить стислу репрезентацію вхідних даних.

– приховані шари декодера: виконують відновлення даних, збільшуючи їхню розмірність.
– вихідний шар: формує відновлені дані, які порівнюються з вхідними.

Простий автоенкодер можна записати як функції:

$$z = f(x) = \partial(W_x + b) \quad (2),$$

$$x = g(z) = \partial(W`_z + b`) \quad (3).$$

де W,W ′ – ваги мережі, b, b′ – зміщення, ∂ – активаційна функція. Якщо W=W′, такий автоенкодер називається симетричним, що спрощує навчання.

Лінійні автоенкодери працюють аналогічно методу головних компонент (PCA), де енкодер і декодер використовують лише лінійні перетворення. Нелінійні автоенкодери використовують глибші мережі та нелінійні активації, що дозволяє навчати складніші представлення.

Варіаційні автоенкодери (VAE) є ймовірнісною версією автоенкодерів, де латентний простір моделюється за допомогою розподілу ймовірностей, зазвичай нормального. Це дозволяє отримати більш узагальнене представлення даних.

Функція втрат для VAE включає не тільки похибку реконструкції, але й регуляризаційний член, який змушує латентні змінні набувати нормального розподілу:

$$L = E_{q(z|x)}[\log p(x|z)] - D_{KL}(q(z|x)||p(z)) \quad (4),$$

де D_{KL} – дивергенція Кульбака-Лейблера між розподілом латентних змінних та нормальним розподілом.

VAE часто використовуються для аномального аналізу в задачах генеративного моделювання.

Шумостійкі автоенкодери (Denoising Autoencoders, DAE): тренуються на відновлення вхідних даних після штучного додавання шуму, що робить їх стійкими до дрібних варіацій даних.

Розріджені автоенкодери (Sparse Autoencoders, SAE): використовують регуляризацію L1 для зменшення кількості активованих нейронів у латентному просторі, що допомагає виявляти ключові особливості нормальних даних.

При застосуванні автоенкодерів для аномального аналізу необхідно визначити поріг для виявлення аномалій, використовуючи один з підходів:

– фіксований поріг: якщо похибка реконструкції перевищує задане значення, зразок вважається аномальним;
– динамічний поріг: визначається на основі статистичних характеристик похибки (наприклад, середнього значення та стандартного відхилення).

$$T = \mu + k\partial \quad (5),$$

де μ – середнє значення похибки реконструкції, ∂ – стандартне відхилення, k – коефіцієнт, що регулює чутливість до аномалій.

Автоенкодери використовуються для виявлення аномалій у різних аспектах мікросервісних архітектур:

– моніторинг метрик продуктивності: аналіз CPU, RAM, мережевого трафіку, кількості запитів тощо;
– аналіз логів: пошук нетипових записів у логах сервісів;
– виявлення відхилень у потоках запитів: аналіз взаємодій між сервісами на предмет незвичних патернів.

Оскільки мікросервіси генерують великі обсяги даних, важливо використовувати автоенкодери з ефективними механізмами обробки потоків даних у реальному часі.

Автоенкодери є потужним інструментом для виявлення аномалій у мікросервісних архітектурах. Вони дозволяють ефективно аналізувати складні багатовимірні дані, автоматично виявляючи аномальні патерни. Проте їх продуктивність залежить від вибору архітектури, метрики втрат та підходу до визначення порогу.

Перспективні напрямки досліджень включають комбінування автоенкодерів із рекурентними нейронними мережами (LSTM) та використання ансамблевих моделей для підвищення точності виявлення аномалій.

2.2. Рекурентні нейронні мережі (RNN) та LSTM

Рекурентні нейронні мережі та їхні модифікації, такі як LSTM (Long Short-Term Memory), ефективно застосовуються для аналізу часових рядів та виявлення аномалій у послідовних даних [6]. На рисунку 2 наведено схематичний принцип роботи LSTM:

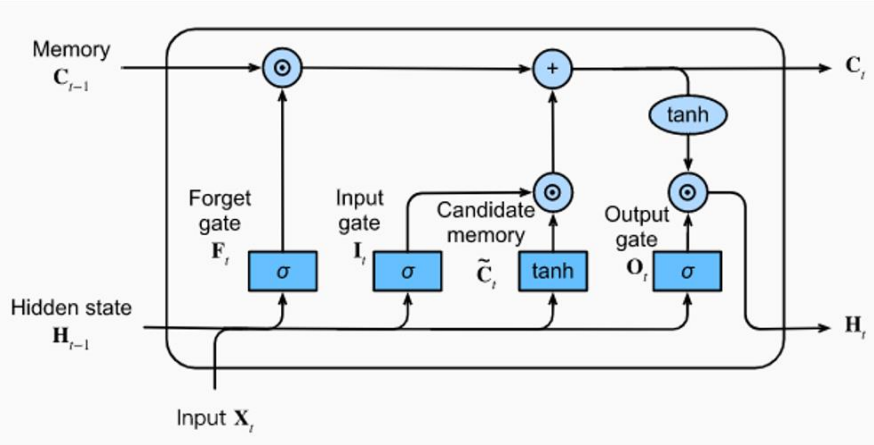

Рис. 2. Схематичний принцип роботи LSTM

Перевага LSTM у порівнянні з класичними RNN полягає у здатності зберігати довготривалі залежності завдяки спеціальним механізмам оновлення стану комірки. Це робить їх ефективними для виявлення відхилень у поведінці мікросервісів, таких як нетипові піки навантаження або відмови у взаємодії сервісів.

На відміну від звичайних багатошарових перцептронів, рекурентні нейронні мережі мають внутрішній стан, який дозволяє зберігати інформацію про попередні кроки. Це досягається за рахунок зворотних зв'язків, що дозволяють передавати вихідні значення одного шару до наступного часу. Завдяки цій властивості RNN добре працюють із послідовними даними, такими як часові ряди або текстові дані.

Математична модель RNN визначається через рекурентне оновлення прихованого стану:

$$h_t = \partial(W_h h_{t-1} + W_x x_t + b) \quad (6),$$

де h_t – прихований стан у момент часу t, W_h та W_x– матриці ваг, x_t – вхідні дані, b – зміщення, а ∂ – функція активації (зазвичай гіперболічний тангенс або сигмоїда).

Однак базові RNN мають обмежену здатність до запам'ятовування довготривалих залежностей. Проблема затухаючих градієнтів ускладнює навчання, що робить такі мережі неефективними для аналізу тривалих послідовностей.

Для розв'язання проблеми затухаючих градієнтів було запропоновано архітектури з механізмом керованої пам'яті, зокрема довготривало-короткочасну пам'ять (LSTM) та керовані рекурентні блоки (GRU).

LSTM-мережі складаються з спеціальних осередків пам'яті, що дозволяють контролювати, яку інформацію зберігати та яку забувати. Кожен LSTM-блок містить три основні елементи: вентиль забування, вентиль входу та вентиль виходу.

Вентиль забування визначає, яку частину попереднього стану слід видалити:

$$f_t = \partial(W_{\text{a}} h_{\text{e}-1} + W_x x_t + b_f) \quad (7)$$

Вентиль входу регулює оновлення нового стану:

$$i_t = \partial(W_i h_{i-1} + W_x x_t + b_i) \quad (8)$$

Кандидат на оновлення стану обчислюється через нелінійну трансформацію:

$$C`_t = tanh(W_c h_{c-1} + W_x x_t + b_c) \quad (9).$$

Поточний стан осередку пам'яті оновлюється як поєднання старої та нової інформації:

$$C_t = f_{t-1} + i_t C`_t \quad (10)$$

Нарешті, вентиль виходу визначає, що саме передавати далі у прихований стан:

$$o_t = \partial(W_o h_{t-1} + W_x x_t + b_o) \quad (11)$$

$$h_t = o_t \tanh(C_t) \quad (12).$$

Ця архітектура дає змогу ефективно моделювати довготривалі залежності та зберігати контекст навіть на великих часових масштабах.

GRU, що є спрощеною версією LSTM, використовує лише два вентилі – оновлення та скидання, що зменшує обчислювальні витрати без значної втрати точності.

Завдяки здатності прогнозувати часові ряди, LSTM широко використовується для виявлення аномалій у поведінці мікросервісів. Основна ідея полягає у тому, щоб навчити модель передбачати наступне значення метрики та порівнювати прогноз із реальним значенням. Якщо різниця між ними перевищує певний поріг, це може свідчити про аномальну поведінку системи.

Під час тренування модель отримує послідовність спостережень, після чого навчається передбачати наступний крок. Наприклад, для аналізу навантаження сервера можна взяти історичні дані про використання CPU та RAM і навчити модель передбачати їх зміну. Якщо спостережувані значення значно відрізняються від прогнозованих, система позначає їх як потенційні аномалії.

Інший підхід полягає у використанні автоенкодерів на основі LSTM. У такій архітектурі енкодер аналізує вхідну послідовність, а декодер намагається відновити її. Чим більше відхилення між вхідними та відновленими значеннями, тим вища ймовірність аномалії.

У реальних мікросервісних архітектурах LSTM використовується для аналізу таких параметрів, як рівень навантаження серверів, зміни в часі відгуку сервісів, аналіз логів та потоків запитів. Наприклад, якщо сервіс зазвичай обробляє запити за 50 мс, а раптово час зростає до 300 мс, це може свідчити про проблему, таку як збій або перевантаження мережі.

Окрім цього, LSTM може бути використана у комбінованих моделях разом із згортковими нейронними мережами (CNN). У такому випадку CNN аналізує короткочасні залежності, а LSTM відповідає за довготривалу динаміку. Цей підхід дозволяє враховувати як локальні, так і глобальні зміни у поведінці системи, підвищуючи точність виявлення аномалій.

Попри високу ефективність, використання LSTM у задачах аномального аналізу має низку викликів. Навчання таких моделей є ресурсомістким процесом, особливо при обробці великих масивів даних у реальному часі. Окрім цього, визначення оптимального розміру вікна послідовності є нетривіальним завданням: занадто коротке вікно може втратити важливий контекст, а надто довге – ускладнити навчання.

Ще однією проблемою є чутливість до вибору гіперпараметрів, таких як кількість шарів, розмір латентного простору та тип функції активації. Неправильна конфігурація може призвести до переобучення або, навпаки, до поганого узагальнення.

LSTM-мережі є потужним інструментом для аналізу часових рядів та виявлення аномалій у мікросервісних системах. Вони дозволяють з високою точністю прогнозувати майбутні стани системи та визначати аномальні відхилення. Однак їхнє використання потребує значних обчислювальних ресурсів, ретельного налаштування параметрів і відповідного підходу до визначення порогів аномальності.

Подальші дослідження можуть бути спрямовані на розвиток гібридних моделей, які комбінують LSTM із автоенкодерами та згортковими мережами, а також на впровадження механізмів оптимізації, що дозволять покращити швидкість роботи без втрати точності.

2.3. Згорткові нейронні мережі (CNN) у задачах аномального аналізу

Згорткові нейронні мережі (CNN) традиційно використовуються у задачах комп'ютерного зору та обробки зображень, оскільки вони ефективно виявляють локальні патерни та структури у вхідних даних. Проте їхні можливості можна застосувати і для аналізу часових рядів та багатовимірних даних у мікросервісних системах. Основна ідея полягає у тому, що згорткові шари можуть виділяти характерні особливості аномальних патернів, незалежно від їхнього розташування у часовій або просторовій послідовності. [7]. На рисунку 3 наведено схематичний принцип роботи CNN.

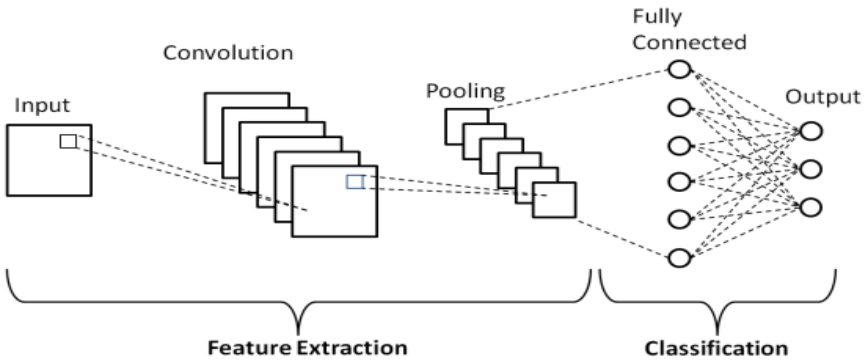

Рис. 3. Схематичний принцип роботи CNN

На відміну від повнозв'язних шарів у класичних нейронних мережах, згорткові шари працюють з локальними областями даних, що дозволяє значно зменшити кількість параметрів та підвищити ефективність обчислень. Для цього використовуються фільтри (ядра згортки), які проходять через вхідні дані та виявляють закономірності у вигляді карт ознак.

Математично операція згортки описується наступним рівнянням:

$$y(i,j) = \sum_m \sum_n x\,(i+m, j+n) w(m,n) \quad (12),$$

де $y(i,j)$ – вхідні дані, w(m,n) – ядро згортки, а y(i,j) – отримане значення після згортки.

Ця операція дозволяє виділяти суттєві характеристики вхідних даних, які можна використовувати для класифікації або аномального аналізу.

Архітектура згорткової нейронної мережі складається з кількох основних компонентів:

- згорткові шари (Convolutional layers) – застосовують фільтри для виділення ознак у вхідних даних;
- шари активації (Activation layers) – додають нелінійність, зазвичай використовується функція ReLU ($ReLU(x) = \max(0, x)$);

- шари підсемплювання (Pooling layers) – зменшують розмірність вхідних даних, зберігаючи найважливішу інформацію (наприклад, max-pooling);
- повнозв'язні шари (Fully Connected layers) – формують остаточне представлення даних для класифікації або оцінки аномальності.

Для задач виявлення аномалій CNN може використовуватися двома способами: як автономна модель, що аналізує багатовимірні вхідні дані, або як частина комбінованої архітектури разом із рекурентними мережами чи автоенкодерами.

Попри те, що CNN спочатку створювали для роботи з двовимірними зображеннями, вони можуть бути адаптовані для аналізу одновимірних часових рядів. Для цього часові дані перетворюються у тривимірний формат, що дозволяє згортковим шарам ефективно обробляти взаємозв'язки між сусідніми часовими точками.

У мікросервісних системах такі моделі можуть використовуватися для моніторингу продуктивності сервісів, аналізу трафіку або оцінки аномальних змін у поведінці системи. Наприклад, якщо серверний вузол зазвичай обробляє певну кількість запитів за секунду, але раптово спостерігається значне відхилення, CNN може виявити подібну зміну як аномальну.

Замість традиційних рекурентних підходів, які аналізують часові ряди послідовно, CNN використовують фільтри для виявлення характерних шаблонів у всій послідовності одночасно. Це дозволяє виявляти складні аномалії, які могли б бути пропущені при традиційному аналізі окремих точок даних.

Хоча згорткові нейронні мережі добре працюють з просторовими залежностями, вони мають певні обмеження у випадку складних часових закономірностей. Тому у практичних застосуваннях CNN часто комбінуються з іншими моделями.

CNN + LSTM – одна з найпопулярніших гібридних архітектур, де згорткові шари відповідають за виділення локальних патернів у часових рядах, а LSTM-мережа аналізує довготривалі залежності. Цей підхід використовується у складних мікросервісних системах, де важливо враховувати як локальні, так і глобальні зміни у поведінці сервісів.

Інший підхід – використання CNN-автоенкодерів. У цьому випадку згорткові шари виконують роль енкодера, стискаючи вхідні дані у компактне представлення, а декодер намагається відновити їх. Висока похибка реконструкції сигналізує про можливу аномалію.

Також CNN можуть бути інтегровані у ансамблеві методи, де результати згорткової мережі комбінуються з іншими моделями (наприклад, деревами рішень або методами кластеризації) для підвищення точності виявлення аномалій.

Згорткові нейронні мережі знаходять застосування у багатьох аспектах моніторингу мікросервісних архітектур:

- аналіз навантаження серверів та розподілу запитів між мікросервісами;
- виявлення атак на рівні мережевого трафіку та аномальної активності користувачів;
- аналіз логів сервісів для пошуку нестандартних подій або помилок;
- автоматичне прогнозування відмов у системах на основі історичних даних.

Одним із важливих факторів у застосуванні CNN є швидкість їхньої роботи. Завдяки паралельному виконанню згорткових операцій CNN можуть аналізувати великі масиви даних у реальному часі, що робить їх корисними для задач оперативного моніторингу.

Попри переваги, CNN мають певні обмеження. Оскільки вони не розраховані на роботу із залежностями у часових рядах, їхня ефективність може бути нижчою у випадках, коли необхідно враховувати довготривалі тренди у даних. Для таких задач рекурентні мережі або їхні гібридні варіанти можуть виявитися більш ефективними.

Ще однією проблемою є необхідність попередньої підготовки даних. Часові ряди часто мають нерівномірні інтервали, шуми та пропущені значення, що може впливати на роботу згорткових моделей. Додаткові етапи попередньої обробки, такі як нормалізація чи інтерполяція даних, можуть бути необхідними для покращення результатів.

Згорткові нейронні мережі є потужним інструментом для виявлення аномалій у мікросервісних системах, особливо у випадках, коли дані мають багатовимірну структуру або містять локальні закономірності. Вони дозволяють швидко обробляти великі масиви інформації та виявляти складні патерни у поведінці системи.

Хоча CNN мають певні обмеження у роботі з часовими рядами, їх поєднання з іншими моделями, такими як LSTM або автоенкодери, дозволяє підвищити точність виявлення аномалій. Подальші дослідження можуть бути спрямовані на оптимізацію гібридних архітектур та використання самонавчальних механізмів для адаптації моделей до нових типів аномалій.

3. Порівняльний аналіз методів

Виявлення аномалій у мікросервісних системах є складним завданням, що потребує вибору оптимального методу залежно від особливостей даних, вимог до продуктивності та точності. Серед найпоширеніших підходів у цій сфері використовуються автоенкодери (AE), рекурентні нейронні мережі (RNN), довготривало-короткочасна пам'ять (LSTM) та згорткові нейронні мережі (CNN).

Автоенкодери є ефективними для виявлення відхилень у багатовимірних даних. Вони навчаються реконструювати нормальні зразки, і якщо похибка реконструкції перевищує певний поріг, об'єкт вважається аномальним. Цей метод працює добре, якщо аномалії значно відрізняються від нормальних даних, проте у випадку поступових змін або аномалій, схожих на нормальні патерни, АЕ можуть не виявити відхилення. До того ж вони не є адаптивними: при зміні поведінки системи автоенкодери потребують перевчання, що є ресурсомістким процесом.

Рекурентні нейронні мережі, зокрема LSTM, враховують часовий контекст, що робить їх корисними для аналізу метрик, що змінюються з часом. Вони ефективно прогнозують майбутні стани системи, порівнюють очікувані та фактичні значення і, у разі значного розходження, позначають спостереження як аномальне. Це робить LSTM одним із найкращих методів для виявлення закономірних відхилень, проте він має високу обчислювальну складність, що ускладнює використання у реальному часі без відповідної оптимізації.

Згорткові нейронні мережі використовуються для аналізу багатовимірних часових рядів, особливо у випадках, коли аномалії проявляються як локальні відхилення у даних. Вони добре працюють завдяки можливості автоматичного виділення особливостей у сигналах, що може бути складним для інших методів. CNN забезпечують високу швидкість обробки завдяки паралельним обчисленням і є придатними для роботи у реальному часі, особливо у разі використання графічних процесорів (GPU). Проте цей підхід менш ефективний для довготривалих часових залежностей, оскільки він аналізує лише локальні патерни, що може призвести до пропуску певних типів аномалій.

З точки зору роботи у реальному часі, найкраще підходять CNN та автоенкодери, оскільки вони швидко обробляють вхідні дані. LSTM мають вищу обчислювальну складність, що може створювати затримки у великих системах. Проте гібридні моделі, наприклад комбінація CNN+LSTM або AE+LSTM, дозволяють досягти кращого балансу між продуктивністю та точністю. Вибір конкретного підходу залежить від типу аномалій, які потрібно виявляти, та вимог до швидкості роботи системи. Далі в таблиці 1 наведено порівняльну характеристику основних методів виявлення аномалій за ключовими параметрами.

Таблиця 1

Порівняльна характеристика основних методів виявлення аномалій

Критерій	Автоенкодери (АЕ)	LSTM	CNN
Точність	Висока, але залежить від якості вибору порогу реконструкції	Висока для часових рядів, добре прогнозує зміни	Висока, якщо аномалії мають характерні локальні патерни
Стійкість до шуму	Помірна, може бути підвищена використанням шумостійких автоенкодерів	Висока, добре фільтрує випадкові відхилення	Висока, оскільки згорткові шари усереднюють нерелевантні деталі
Адаптивність	Низька, потребує перевчання при зміні характеристик системи	Висока, здатність адаптуватися до змін у часових рядах	Помірна, нові шаблони можуть вимагати донавчання
Обчислювальна складність	Помірна, швидка обробка після навчання	Висока, особливо для довгих часових послідовностей	Низька, ефективна робота на GPU
Можливість роботи у реальному часі	Висока після навчання, швидке обчислення похибки	Середня, залежить від складності моделі та довжини послідовності	Висока, добре підходить для потокового аналізу
Чутливість до типу даних	Найкраще працює на стаціонарних наборах даних	Підходить для часових рядів із довготривалими трендами	Найкраще працює на просторових та короткочасних залежностях
Складність налаштування	Середня, необхідний підбір порога аномалій	Висока, потребує великої кількості гіперпараметрів	Низька, оскільки стандартні архітектури працюють добре
Придатність до великих даних	Висока, обчислення реконструкції масштабуються добре	Середня, складність зростає з довжиною послідовностей	Висока, особливо при використанні GPU

Загалом, ефективність кожного підходу визначається специфікою даних та вимогами до продуктивності. Автоенкодери добре підходять для виявлення загальних аномалій у стаціонарних наборах даних, але потребують повторного навчання у разі змін у поведінці системи. LSTM забезпечують високу точність при аналізі часових рядів, але вимагають значних обчислювальних

ресурсів, що може ускладнювати їх використання у реальному часі. CNN працюють швидко та ефективно, особливо при потоковому аналізі багатовимірних даних, але можуть не враховувати довготривалі залежності, якщо вони не представлені у локальних шаблонах.

Комбіновані підходи, такі як AE+LSTM або CNN+LSTM, можуть компенсувати слабкі сторони окремих моделей і забезпечити кращу загальну продуктивність. Наприклад, згорткові шари можуть швидко виділяти основні особливості у часових рядах, після чого LSTM-мережа аналізуватиме довготривалі взаємозв'язки між ними. Так само AE можуть використовуватися для попереднього зменшення розмірності вхідних даних перед обробкою їх LSTM, що знижує обчислювальне навантаження.

Вибір конкретного методу залежить від того, що є критично важливим у системі моніторингу: швидкість обробки чи точність виявлення складних аномалій. Якщо потрібна швидка оцінка великих потоків даних у реальному часі, CNN будуть найкращим варіантом. Для виявлення складних закономірностей у часових рядах найефективнішими будуть LSTM. Автоенкодери є хорошим компромісним варіантом для загального виявлення відхилень, але потребують ретельного налаштування параметрів порогу аномалій.

З урахуванням усіх аспектів, використання гібридних підходів залишається найбільш перспективним напрямком розвитку методів виявлення аномалій у мікросервісних системах. Подальші дослідження можуть бути спрямовані на створення адаптивних ансамблевих моделей, що комбінують переваги кожного підходу та здатні самостійно оновлюватися у процесі роботи [8].

Висновки

Дослідження методів виявлення аномалій у мікросервісних системах показало, що нейронні мережі є ефективним інструментом для аналізу складних та високорозмірних даних. Автоенкодери, рекурентні нейронні мережі (LSTM) та згорткові нейронні мережі (CNN) мають свої переваги та недоліки, що впливають на їхню придатність до використання у різних сценаріях.

Автоенкодери показують високу ефективність при виявленні загальних аномалій, оскільки навчаються реконструювати лише нормальні патерни. Висока похибка реконструкції є ключовим критерієм для визначення аномальних випадків, що робить цей метод корисним для роботи зі складними багатовимірними даними. Проте їхнім основним недоліком є низька адаптивність, оскільки при зміні характеристик системи необхідно перевчати модель. Крім того, автоенкодери можуть неефективно працювати у випадках, коли аномалії лише незначно відрізняються від нормальних зразків.

Рекурентні нейронні мережі, зокрема LSTM, демонструють високу точність у випадках, коли важливо враховувати часову структуру даних. Вони здатні прогнозувати майбутні значення метрик мікросервісів, що дозволяє їм виявляти відхилення від нормальної поведінки. Їхньою перевагою є здатність аналізувати довготривалі залежності, що важливо для аналізу трендів у роботі мікросервісів. Однак висока обчислювальна складність та вимоги до ресурсів обмежують їхнє використання у реальному часі без оптимізації архітектури або зменшення обсягу оброблюваних даних.

Згорткові нейронні мережі є оптимальним вибором для задач, де аномалії проявляються у вигляді локальних патернів у даних. Вони забезпечують швидку та ефективну обробку великих обсягів інформації, особливо при використанні апаратного прискорення, такого як GPU. CNN добре підходять для аналізу короткотривалих залежностей, але вони менш ефективні для довготривалих змін у часових рядах, що може призводити до пропуску деяких видів аномалій.

Порівняння методів показало, що кожен з них має обмеження, які необхідно враховувати при виборі підходу для конкретного застосування. У реальних мікросервісних системах часто доцільно використовувати комбіновані моделі, що поєднують переваги різних підходів. Наприклад, комбінація CNN та LSTM дозволяє одночасно виявляти локальні відхилення та аналізувати довготривалі залежності, що забезпечує кращу загальну продуктивність. Автоенкодери у поєднанні з LSTM можуть зменшити обчислювальне навантаження, спочатку стискаючи вхідні дані перед їх подальшим аналізом.

Подальші дослідження у цій галузі можуть бути спрямовані на покращення адаптивності моделей до змін у поведінці мікросервісних систем. Одним із перспективних напрямків є розробка самонавчальних нейронних мереж, які можуть оновлювати свої параметри в процесі роботи без необхідності повного перевчання. Це дозволить значно підвищити точність виявлення аномалій у динамічних середовищах, де параметри системи змінюються у часі. Ще одним важливим напрямком є створення гібридних ансамблевих моделей, які комбінують різні нейронні архітектури та традиційні методи машинного навчання. Поєднання нейронних мереж із методами кластеризації або деревами рішень може забезпечити більшу стійкість до шуму та покращити інтерпретованість отриманих результатів. Також доцільно дослідити вплив трансформерних архітектур на задачу виявлення аномалій у мікросервісних системах. Оскільки трансформери демонструють високу ефективність у моделюванні довготривалих залежностей у текстових і часових даних, вони можуть стати конкурентоспроможною альтернативою LSTM, особливо у випадках, коли необхідно швидко обробляти великі обсяги послідовних даних.

Крім того, перспективним є розробка алгоритмів для зменшення обчислювальної складності, які дозволять використовувати нейронні мережі у реальному часі без значних ресурсних витрат. Це

може включати техніки прунінгу (видалення непотрібних нейронів), квантизації ваг та оптимізації архітектури шляхом використання глибших, але менш широких шарів.

Окремим важливим питанням залишається автоматизоване налаштування моделей, оскільки вибір гіперпараметрів значно впливає на продуктивність алгоритмів. Використання методів автоматичного пошуку архітектур (NAS – Neural Architecture Search) може допомогти знайти оптимальні конфігурації моделей для конкретних задач виявлення аномалій.

У цілому, подальші дослідження мають бути зосереджені на створенні адаптивних, обчислювально ефективних і масштабованих методів, що здатні працювати у реальному часі та зменшувати кількість хибнопозитивних спрацьовувань. Це дозволить розширити можливості моніторингу мікросервісних систем і зробити їх більш стійкими до збоїв та атак.

Література

1. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys, 41*(3), 1–58. https://doi.org/10.1145/1541880.1541882
2. Pimentel, M. A. F., Clifton, D. A., Tarassenko, L., & et al. (2014). A review of novelty detection. *Signal Processing, 99*, 215–249. https://doi.org/10.1016/j.sigpro.2013.12.026
3. Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, June). LOF: Identifying density-based local outliers. In *Proceedings of the ACM SIGMOD International Conference on Management of Data* (pp. 93–104). https://doi.org/10.1145/335191.335388
4. Liu, F. T., Ting, K. M., & Zhou, Z. (2008, December). Isolation Forest. In *Proceedings of the 8th IEEE International Conference on Data Mining* (pp. 413–422). https://doi.org/10.1109/ICDM.2008.17
5. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. https://www.deeplearningbook.org
6. Chalapathy, R., & Chawla, S. (2019). Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*. https://arxiv.org/abs/1901.03407
7. Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science, 313*(5786), 504–507. https://doi.org/10.1126/science.1127647
8. Malhotra, P., Vig, L., Shroff, G., & Agarwal, P. (2015, April). Long short-term memory networks for anomaly detection in time series. In *Proceedings of the 23rd European Symposium on Artificial Neural Networks (ESANN)* (pp. 89–94). https://arxiv.org/abs/1502.06690
9. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
10. Pascanu, R., Mikolov, T., & Bengio, Y. (2013, June). On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning (ICML)* (pp. 1310–1318). https://arxiv.org/abs/1211.5063
11. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539
12. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)* (pp. 1097–1105). https://doi.org/10.1145/3065386
13. Ruff, L., Vandermeulen, R. A., Görnitz, N., et al. (2018). Deep one-class classification. In *Proceedings of the 35th International Conference on Machine Learning (ICML)* (pp. 4390–4399). https://arxiv.org/abs/1801.05376
14. Zong, B., Song, Q., Min, M. R., et al. (2018). Deep autoencoding Gaussian mixture model for unsupervised anomaly detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1802.00187
15. Xu, H., Chen, Z., Liu, J., et al. (2018, April). Unsupervised anomaly detection via variational auto-encoder for seasonal KPIs in web applications. In *Proceedings of the 27th International Conference on World Wide Web (WWW)* (pp. 187–196). https://doi.org/10.1145/3178876.3185996.

References

1. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys, 41(3), 1–58. https://doi.org/10.1145/1541880.1541882
2. Pimentel, M. A. F., Clifton, D. A., Tarassenko, L., & et al. (2014). A review of novelty detection. Signal *Processing, 99*, 215–249. https://doi.org/10.1016/j.sigpro.2013.12.026
3. Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, June). LOF: Identifying density-based local outliers. In *Proceedings of the ACM SIGMOD International Conference on Management of Data* (pp. 93–104). https://doi.org/10.1145/335191.335388
4. Liu, F. T., Ting, K. M., & Zhou, Z. (2008, December). Isolation Forest. In *Proceedings of the 8th IEEE International Conference on Data Mining* (pp. 413–422). https://doi.org/10.1109/ICDM.2008.17
5. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. https://www.deeplearningbook.org
6. Chalapathy, R., & Chawla, S. (2019). Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*. https://arxiv.org/abs/1901.03407

7. Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science, 313*(5786), 504–507. https://doi.org/10.1126/science.1127647

8. Malhotra, P., Vig, L., Shroff, G., & Agarwal, P. (2015, April). Long short-term memory networks for anomaly detection in time series. In *Proceedings of the 23rd European Symposium on Artificial Neural Networks (ESANN)* (pp. 89–94). https://arxiv.org/abs/1502.06690

9. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

10. Pascanu, R., Mikolov, T., & Bengio, Y. (2013, June). On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning (ICML)* (pp. 1310–1318). https://arxiv.org/abs/1211.5063

11. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

12. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)* (pp. 1097–1105). https://doi.org/10.1145/3065386

13. Ruff, L., Vandermeulen, R. A., Görnitz, N., et al. (2018). Deep one-class classification. In *Proceedings of the 35th International Conference on Machine Learning (ICML)* (pp. 4390–4399). https://arxiv.org/abs/1801.05376

14. Zong, B., Song, Q., Min, M. R., et al. (2018). Deep autoencoding Gaussian mixture model for unsupervised anomaly detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1802.00187

15. Xu, H., Chen, Z., Liu, J., et al. (2018, April). Unsupervised anomaly detection via variational auto-encoder for seasonal KPIs in web applications. In *Proceedings of the 27th International Conference on World Wide Web (WWW)* (pp. 187–196). https://doi.org/10.1145/3178876.3185996