

# СТВОРЕННЯ СИНТЕТИЧНОГО НАБОРУ ДАНИХ ДЛЯ ОЦІНЮВАННЯ АРХІТЕКТУР НЕЙРОМЕРЕЖЕВИХ МОДЕЛЕЙ

<sup>1</sup> Вінницький національний технічний університет

## Анотація

Розробка моделей глибокого навчання для обробки часово-залежних послідовностей, таких як відео, базові часові ряди, або сигнали робототехнічних систем, є складною задачею, особливо на етапах початкової оцінки ефективності розроблених архітектур. Використання реальних датасетів, що характеризуються високою стохастичністю та складністю, ускладнює ізольований аналіз властивостей архітектур та робить вартісним обчислювальний процес. У роботі запропоновано новий підхід до створення штучного набору даних, який генерує нескінченний потік унікальних послідовностей з 2D або 3D фігур. Цей датасет призначений для створення контрольованого середовища базової оцінки та порівняння архітектур моделей глибокого навчання перед їх застосуванням до складніших реальних задач, забезпечуючи можливість ізольованого вивчення ключових характеристик моделей.

**Ключові слова:** штучний інтелект, машинне навчання, оцінка архітектур, синтетичні дані.

## Abstract

The development of deep learning models for processing time-dependent sequences, such as video, time series, or robotics signals, faces the challenge of evaluating the effectiveness of new architectures in the early stages of development. The use of real-world datasets, characterized by high stochasticity and complexity, complicates the isolated study of architecture properties and requires significant computational resources. A novel synthetic dataset proposed that allows for the generation of an infinite stream of unique sequences of 2D or 3D figure. The dataset is designed to create a controlled environment for basic evaluation and comparison of deep learning model architectures before their application to more complex real-world tasks, ensuring possibility of isolated study of key model characteristics.

**Keywords:** artificial intelligence, machine learning, architecture evaluation, synthetic data.

Розробка та валідація моделей для обробки часово-залежних послідовностей є ключовою задачею в сфері глибокого навчання, що охоплює широкий спектр застосувань, від аналізу відео та прогнозування часових рядів до керування робототехнічними системами. Для навчання таких моделей використовуються різноманітні датасети з реальними даними. Серед них відомі Kinetics-400 [1] для відео аналізу, датасети для прогнозування часових рядів, як, наприклад, Electricity Transformer Dataset (ETT) [2], та датасети сигналів з робототехнічних систем, такі як Robotcar [3]. Ці датасети є цінними для навчання та оцінки моделей, проте, на етапі розробки нових архітектур, їх використання може бути контрпродуктивним через високу стохастичність, багатовимірність та значні обчислювальні витрати. Для початкової перевірки концепцій та базової оцінки архітектурних рішень необхідне спрощене, контрольоване середовище. Таке середовище має дозволяти ізолювати ключові властивості моделей та забезпечувати стабільність і відтворюваність експериментів, незалежно від стохастичної природи реальних даних.

Для спрощення базового тестування архітектур, що працюють з часово-залежними послідовностями, вже існують певні рішення. Прикладами є датасет обертових цифр MNIST [4], де цифри з класичного датасету MNIST піддаються обертанню, та Moving MNIST [5], де цифри рухаються у випадковому напрямку по сцені. Ці датасети дозволяють тестувати здатність моделей до розпізнавання об'єктів при зміні їхньої орієнтації або положення, і є корисними для вирішення вузького кола задач. Проте, їх варіативність залишається обмеженою, що може бути недостатнім для глибокого розуміння властивостей архітектур щодо передбачення у часових послідовностях. Основний акцент в цих датасетах робиться на розпізнаванні цифр, а не на моделюванні часової динаміки як такої. Крім того, фіксований та обмежений розмір цих датасетів може призвести до

перенавчання моделей та ускладнює оцінку їхньої здатності до узагальнення.

Для вирішення зазначених проблем, у роботі розроблено синтетичний датасет, спеціально призначений для базової оцінки архітектур моделей, що застосовуються для часово-залежних послідовностей [6]. Основна ідея полягає в генерації послідовностей руху простих геометричних фігур за чітко визначеними правилами, що дозволяє створювати набори даних з високою передбачуваністю. Датасет підтримує генерацію послідовностей як для простих 2D фігур, таких як трикутники та квадрати (приклад зображено на рисунку 1), так і для 3D фігур, скомпонованих зі з'єднаних по граням кубів (приклад зображено на рисунку 2).

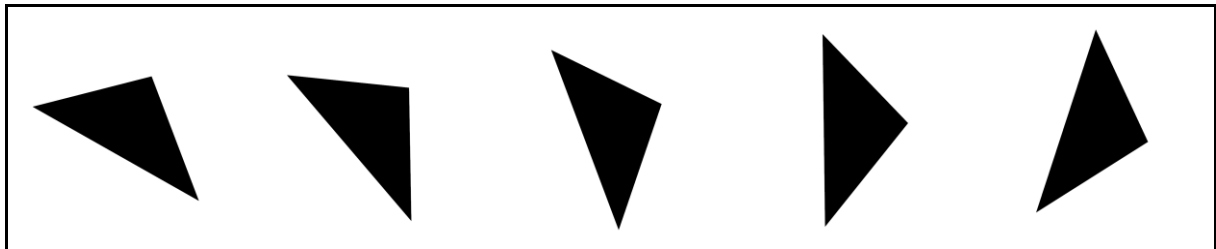


Рисунок 1 – Приклад згенерованої послідовності для 2D фігури

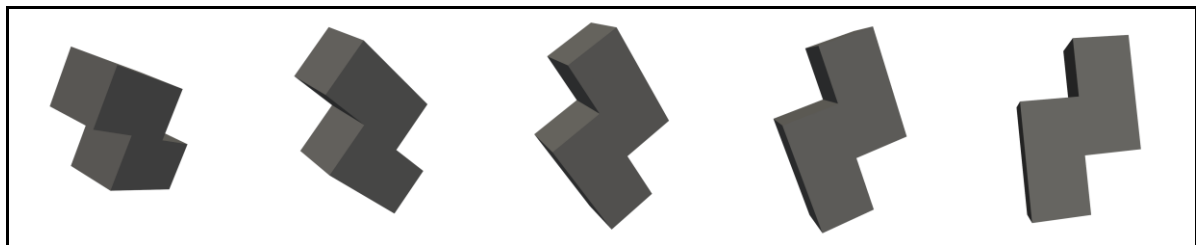


Рисунок 2 – Приклад згенерованої послідовності для 3D фігури

Ключові характеристики датасету включають:

- можливість вибору та налаштування різних типів руху, включаючи постійну швидкість, прискорення/уповільнення, осциляцію, переривання та інші варіації;
- процедурна генерація кожного елементу послідовності, що забезпечує нескінченний потік унікальних прикладів для уникнення перенавчання та масштабованість тестування;
- низька ентропія для базового тестування: особливо в 2D конфігурації, датасет забезпечує низьку ентропію, що є ідеальним для базової перевірки працездатності архітектур та їх ключових властивостей, таких як здатність до запам'ятовування послідовностей та ефективність обробки часових залежностей;
- можливість налаштування складності 3D фігур, що дозволяє розширити спектр тестування та оцінити стійкість архітектур до зростання складності вхідних даних;
- збереження детальних метаданих, що фіксують тип фігури, параметри руху, та інші налаштування сцени, забезпечуючи відтворюваність експериментів.

Переваги запропонованого датасету для базової оцінки архітектур полягають у створенні контрольованого та простого середовища, що є оптимальним для початкової перевірки концепцій та архітектур моделей для часово-залежних послідовностей. Розроблений підхід дозволяє оцінити різні аспекти ефективності архітектур, включаючи ефективність використання пам'яті, оскільки створення нових елементів може відбуватись у вигляді нескінченного потоку, не вимагаючи значних ресурсів пам'яті для збереження та завантаження даних.

## Висновки

Запропоновано синтетичний датасет, який є цінним для дослідників та розробників моделей глибокого навчання, що працюють з часово-залежними послідовностями. Він забезпечує контрольоване, відтворюване та масштабоване середовище для базової оцінки архітектур, дозволяючи зосередитися на вивченні фундаментальних властивостей моделей перед їх застосуванням до складних реальних задач. Акцент на створенні датасетів з низькою ентропією для

2D конфігурації робить його особливо привабливим для фундаментальних досліджень та аналізу ефективності архітектур, призначених для обробки часово-залежних послідовностей.

## СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Will Kay, Joao Carreira, et al “The Kinetics Human Action Video Dataset” arXiv:1705.06950 [cs.CV], May 2017.
2. Haoyi Zhou, Shanghan Zhang, et al “Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting” arXiv:2012.07436 [cs.LG], March 2021.
3. Will Maddern, Geoffrey Pascoe, et al “Real-time Kinematic Ground Truth for the Oxford RobotCar Dataset” arXiv: 2002.10152 [cs.RO], Feb. 2020.
4. Shumeet Baluja “Adding Conditional Control to Text-to-Image Diffusion Models” Advances in Neural Information Processing Systems, 11, NIPS 1998.
5. Nitish Srivastava, Elman Mansimov, Ruslan Salakhutdinov “Unsupervised Learning of Video Representations using LSTMs” arXiv: 1502.04681 [cs.LG], Jan. 2016.
6. Simple Shape Dataset Toolbox [Електронний ресурс]. — Режим доступу: <https://github.com/leo27heady/simple-shape-dataset-toolbox> (дата звернення: 20.03.2025). — Назва з екрана.

**Кулик Леонід Русланович** – аспірант факультету інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: leonidkulik2707@gmail.com

**Мокін Олександр Борисович** – доктор технічних наук, професор, професор кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, Вінниця, e-mail: abmokin@vntu.edu.ua

**Kulyk Leonid** – a graduate student, Faculty of Intelligent Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, email: leonidkulik2707@gmail.com

**Mokin Oleksandr** – Dr. Sc. (Eng.), Professor of the Department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: abmokin@vntu.edu.ua