УДК 004.8

ПОРІВНЯЛЬНИЙ АНАЛІЗ АРХІТЕКТУР TRANSFORMER ТА MAMBA В ЗАДАЧАХ ОБРОБКИ ПРИРОДНОЇ МОВИ

Єрьоменко М.І., Губін В.О.
e-mail: maksym.ieromenko@nure.ua, e-mail: vadim.gubin@nure.ua
Харківський національний університет радіоелектроніки, каф. ШІ
м. Харків, Україна

This work is devoted to a comparative analysis of two prominent neural network architectures, Transformer and Mamba, in the context of natural language processing tasks. The study examines the fundamental limitations of the Transformer model, particularly its quadratic computational complexity O(L^2) regarding sequence length, which poses significant challenges for processing long-range dependencies. In contrast, the newly emerged Mamba architecture, based on Selective State Space Models (SSM), is analyzed for its ability to achieve linear scaling O(L) and superior inference speed. The research provides an empirical evaluation of both models across various benchmarks, highlighting their efficiency in memory consumption and text generation quality. Special attention is paid to the trade-offs between the established attention mechanisms and the innovative selective recurrence approach. The findings suggest that while Transformers remain a standard for complex reasoning, Mamba offers a promising and hardware-efficient alternative for future large-scale language modeling.

На сучасному етапі розвитку інтелектуальних інформаційних систем архітектура Transformer є фундаментальною основою [1] для побудови великих мовних моделей. Ефективність таких моделей під час розв’язання задач з обробки природної мови (NLP) зумовлена використанням механізму самоуваги (Self-Attention) [1], що дозволяє моделювати глобальні залежності в даних. Проте, із розширенням сфери застосування ШІ на аналіз наддовгих послідовностей (юридичні документи, програмний код, геномні дані), стає критичною проблема квадратичної обчислювальної складності Transformer-моделей [1], що особливо проявляється у задачах обробки довгих сутностей та послідовностей у NLP, зокрема в задачах розпізнавання іменованих сутностей [2]. Актуальним напрямом досліджень є пошук альтернативних архітектур, серед яких найбільш перспективною є модель Mamba [3], заснована на селективних просторах станів.

Метою роботи є проведення порівняльного аналізу двох архітектур Transformer та Mamba за критеріями обчислювальної складності, швидкості генерації та якості контекстуального моделювання в задачах NLP.

У класичній архітектурі Transformer кожен токен послідовності взаємодіє з кожним іншим токеном через обчислення матриць Q, K та V [1], рисунок 1.

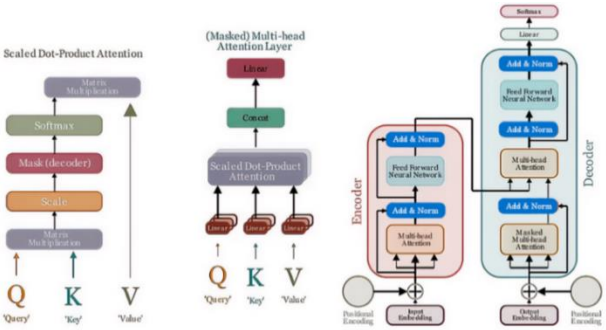

Рисунок 1 – Загальна архітектура моделі Transformer
та механізми Scaled Dot-Product Attention та Multi-Head Attention

Обчислювальна складність цього процесу визначається як [1]:

$$C = O(L^2 \cdot d),$$

де L – довжина послідовності, d – розмірність прихованого стану.

При збільшенні вхідного контексту обсяг необхідної пам'яті для зберігання Key-Value кешу (KV-cache) зростає лінійно, що створює «вузьке місце» для апаратних ресурсів під час інференсу (процесу генерації).

Архітектура Mamba базується на модернізованих моделях простору станів (SSM). Ключовою інновацією є впровадження механізму селективності (S6 mechanism, рисунок 2), який дозволяє параметрам моделі бути функціями від вхідних даних [3].

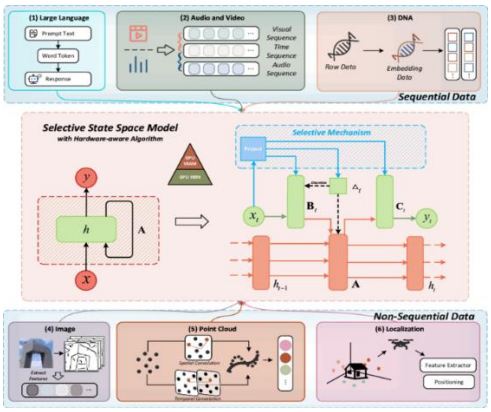

Рисунок 2 – Архітектурна схема блоку Mamba та механізм селективного простору станів (S6)

Це забезпечує моделі здатність до стиснення контексту в стан фіксованого розміру, ігноруючи несуттєву інформацію. Обчислювальна складність Mamba є лінійною [3]:

$$C = O(L \cdot d).$$

Це дозволяє моделі обробляти послідовності довжиною в мільйони токенів без квадратичного уповільнення обчислень.

За результатами аналізу сучасних емпіричних досліджень встановлено, що Mamba демонструє вищу пропускну здатність (throughput) при генерації тексту – від 4 до 5 разів [4] порівняно з Transformer при однаковій кількості параметрів. Mamba є особливо ефективною у задачах, де важлива швидкість обробки в реальному часі та енергоефективність. Однак, Transformer зберігає перевагу в завданнях In-context learning [1], де необхідно вилучати специфічні факти з контексту, оскільки механізм Attention надає прямий доступ до всіх попередніх токенів, тоді як Mamba покладається на стиснутий стан [3]. У таблиці 1 наведені основні відмінності цих архітектур.

Таблиця 1 – Відмінності архітектур Transformer та Mamba

Характеристика	Transformer	Mamba (SSM)
Обчислювальна складність	$O(L^2)$	$O(L)$
Робота з довгим контекстом	Обмежена KV-кешем	Ефективна (стиснутий стан)
Швидкість генерації	Базова	У 4–5 разів вища
In-context learning	Висока якість	Потребує вдосконалення

Порівняльний аналіз підтвердив, що архітектура Mamba є технологічним проривом [2, 3я] у подоланні проблеми «довгого контексту». Вона забезпечує лінійне масштабування та високу швидкість інференсу, що робить її конкурентоспроможною для впровадження в мобільні та розподілені інтелектуальні системи. У той же час, Transformer залишається еталоном [1] для складних аналітичних задач. Перспективою подальших досліджень є розробка гібридних архітектур, що інтегрують блоки Self-Attention у структуру Mamba для балансу між якістю та швидкістю обробки.

Список використаних джерел:

1. Vaswani A. et al. Attention Is All You Need. arXiv.org e-Print archive. 2017. URL: https://arxiv.org/pdf/1706.03762 (дата звернення: 19.02.2026).

2. Shatalov O., Ryabova N. Named Entity Recognition Problem for Long Entities in English Texts. Home Page. DOI: https://doi.org/10.1109/CSIT52700.2021.9648768.

3. Gu A., Dao T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv.org e-Print archive. 2023. URL: https://arxiv.org/pdf/2312.00752 (дата звернення: 19.02.2026).

4. Moonlight Project, NVIDIA, Cornell. An Empirical Study of Mamba-based Language Models. *arXiv.org e-Print archive*. 2024. URL: https://arxiv.org/pdf/2406.07887 (дата звернення: 19.02.2026).