

*В.О. Лесик, А.Ю. Дорошенко*

## КЕРОВАНЕ ДАНИМИ СТИСНЕННЯ ІНФОРМАЦІЇ: ОЦІНКА МЕТОДІВ НА ОСНОВІ МАШИННОГО НАВЧАННЯ

Новітні методи та системи стиснення даних пройшли значний шлях від появи нових методів штучного інтелекту до широкого застосування методів різного роду та призначення, що переповнило простір дослідження та ускладнило процес відбору та пошуку кандидатів моделей для ефективного застосування та подальшого покращення і дослідження. У цій статті запропоновано емпіричний підхід порівняння методів та моделей навченого стиснення у окремих рамках, де виділено потенційну галузь застосування, найкращі моделі на даний момент розвитку, особливості, переваги та недоліки поточних моделей. Проведено оцінку загального застосування навченого стиснення та теоретичних лімітів ентропії. Виконано ряд порівнянь із врахуванням різних предметних областей застосування моделей, описано ряд критеріїв для порівняння та оцінки ефективності імплементації кожного з розглянутих методів, виділено основні проблемні області, які потрібно оцінити перед імплементацією моделей у технологічні процеси, використанням з точки зору звичайних користувачів програмних застосунків чи пакетів для стиснення даних. Надано додаткову оцінку потенційних технологій для подальшого розвитку та покращення методів навченого стиснення, разом з окресленням систем та технологій, що втрачають потенціал до прогресу. Представлений аналіз дозволяє дослідникам отримати загальне уявлення сучасного стану дослідження використання машинного навчання в проблемах стиснення різного роду, оцінити стан розвитку технологій та на основі наведених обмежень та проблемних областей обрати необхідний вектор для подальшого проведення досліджень або пошуку наукової літератури. Ключові слова: машинне навчання, навчене стиснення даних, нейромережеве стиснення даних, гібридні методи стиснення, штучний інтелект.

*V.O. Lesyk, A. Y. Doroshenko*

## DATA-DRIVEN INFORMATION COMPRESSION: AN EVALUATION BASED ON MACHINE LEARNING

The latest data compression methods and systems have come a long way, from the emergence of new artificial intelligence techniques to the widespread use of methods of various types and purposes, which has flooded the research landscape and complicated the process of selecting and identifying candidate models for effective application, further improvement, and research. This article proposes an empirical approach to comparing trained compression methods and models within specific frameworks, highlighting potential application areas, the best models at the current stage of development, and the characteristics, advantages, and disadvantages of current models. An assessment of the general applicability of trained compression and the theoretical limits of entropy is conducted. A series of comparisons is performed, taking into account various subject areas of model application; a set of criteria for comparing and evaluating the implementation effectiveness of each of the considered methods is described; key problem areas that need to be assessed prior to implementing the models into technological processes and as well as their use from the perspective of ordinary users of software applications or data compression packages. An additional assessment of potential technologies for the further development and improvement of learned compression methods is provided, along with an identification of systems and technologies that are losing their potential for progress. This analysis allows researchers to gain a general understanding of the current state of research on the application of machine learning to various compression problems, assess the state of technological development, and—based on the limitations and problem areas identified—select the appropriate direction for further research or for scientific literature discovery. Keywords: machine learning, learned data compression, data compression with neural networks, hybrid compression methods, artificial intelligence.

### Вступ

Традиційні методи стиснення десятиліттями слугували основою інфраструктури цифрових медіа. Ці класичні підходи

спираються на розроблені «вручну» алгоритми, засновані на кодуванні перетворення, квантуванні та ентропійному коду-

ванні, розроблені протягом десятиліть досліджень обробки сигналів та стандартизовані міжнародними організаціями.

На відміну від традиційних методів, що залежать від фіксованих математичних перетворень та евристичної оптимізації, системи стиснення на основі машинного навчання вивчають оптимальні представлення безпосередньо з даних, адаптуючи свої стратегії кодування до статистичних властивостей даних та контенту, що стискається. Цей перехід від стиснення на основі правил до стиснення та на основі даних являє собою не просто поступове покращення, а потенційну зміну парадигми в тому, як ми підходимо до фундаментального бачення стиснення загалом, зміщуючи увагу з ентропійної складової даних до вивчення внутрішньої структури даних та правил, що насамперед будують ці дані. Зазвичай, стиснення досягається за допомогою відкидання перцептивно нерелевантної інформації, часто досягаючи кращої продуктивності порівняно з традиційними методами стиснення. Але новітні застосування нейронних мереж та машинного навчання показують значну перевагу як у стисненні із втратами, так і у стисненні без втрат, що асоціюється із здатністю нейронних моделей імітувати процес біологічного розуміння поданої інформації[1,2].

З точки зору історичного розвитку навчене стиснення — це стиснення даних за допомогою нейронних мереж: моделі навчаються на основі даних, як перетворювати зображення, відео або інші сигнали на компактні латентні коди та відтворювати їх, оптимізуючи водночас компроміс між швидкістю передачі даних та спотворенням[3]. Навчене стиснення розширює можливості класичного кодування за допомогою перетворення, замінюючи модулі, розроблені вручну, на нелінійні моделі, що піддаються навчанню, — зазвичай CNN, RNN, VAE або новіші компоненти на основі трансформерів[3]. У широкому сенсі, навчене стиснення не обмежується лише елементами нейронних мереж, а може бути асоційованим із будь-яким застосуванням машинного навчання, штучного інтелекту, великих мовних моделей чи інтелектуаль-

них методів, що отримують моделі стиснення безпосередньо з даних[3].

Наукові дослідження в галузі стиснення даних беруть свій початок у кінці 1980-х та 1990-х років, коли нейронні мережі вперше почали застосовувати для стиснення зображень, однак ці ранні підходи не були практично реалізовані через обмежені обчислювальні ресурси та масштаб моделей. У 1990-х роках досліджувалися методи стиснення на основі багатопарового перцептрона та випадкових нейронних мереж, але вони так і залишилися непрактичними[4]. Явний перехід від класичного до навчального стиснення відбувся 2015 року, коли Тодерічі та ін. застосували рекурентні нейронні мережі для навчання методів стиснення зображень, що стало першою широко визнаною архітектурою наскрізного («end-to-end») нейронного стиснення[3, 5]. За 10 років подальшого розвитку значний прогрес у галузі навченого стиснення демонструє потенціал технологій машинного навчання, проте досі залишається чимало проблем, що потребують дослідження, а це вимагає проведення систематичного порівняльного аналізу для виявлення ключових напрямків, у яких можна вдосконалити наскрізні та інші архітектури[5, 6].

Попри успіхи в науковій сфері, впровадження методів навчання стиснення в промисловості практично не відбулося; деякі дослідники обрали альтернативний підхід, використовуючи глибоке навчання для підвищення точності стиснення класичних алгоритмів, а не для їх повної заміни[7].

Ландшафт дослідження навченого стиснення за декілька років став перенасичений та занадто спеціалізований, де отримання нових результатів та покращень стає надзвичайно складним та ресурсомістким процесом. Наявні дані підтверджують явища перенасичення, спеціалізації та зростання складності входу, але частіше свідчать про зменшення практичного запасу можливостей та компроміси, пов'язані зі складністю, ніж про повне вичерпання інноваційного потенціалу[7]. Виникає потреба у новій методології аналізу та оцінки даних моделей для виділення найкращих

рішень для імплементції в робочих процесах, разом із пришвидшенням пошуку релевантної інформації про новітні системи стиснення для вибору подальшого руху дослідження та покращення систем і методів стиснення даних.

У даній статі запропоновано методологію спрощеного порівняння існуючих моделей, їхньої релевантності та ефективності, порушено питання меж ефективності застосування навченого стиснення та оцінки методів, що не підпадають під загальну модель класифікації.

### **Аналіз та оцінка методів навченого стиснення та загальна класифікація**

Оскільки стиснення наближається до теоретичних меж, подальше вдосконалення за допомогою класичних методів стає дедалі складнішим. Обмеження класичного стиснення впливають головним чином з їхньої обмеженої здатності моделювати складні розподіли даних. Це мотивує використання машинного навчання, зокрема нейронних мереж, які пропонують вивчені представлення, покращене моделювання контексту та загальну адаптивність. Методи навченого стиснення є видозміненими, але все ще побудовані на спільних принципах із класичними методами: як правило, існують кодер, декодер, стиснене представлення даних та власне методи, що дозволяють знайти чи видобути дане представлення. Навчені моделі стиснення можна класифікувати за тією самою структурною схемою, що й класичні методи — кодування перетворенням, квантування та ентропійне кодування, — оскільки вони успадковують ці етапи, але замінюють модулі, розроблені вручну, нейронними мережами, оптимізованими під дані[3]. Загальне застосування машинного навчання та інтелектуальних методів для стиснення класифікується ідентично до класичних методів: поділ на стиснення без втрат та із втратами, специфікація щодо типу даних — зображення, текстові, відео, аудіодані, послідовності даних та наукові бази. Оцінка та аналіз спираються на такі ж параметри, що і класичні методи, —

коефіцієнт стиснення, швидкість кодування та декодування, вага кодера та декодера, необхідна пам'ять. Навчене стиснення розширює дані параметри, але потребує додаткову увагу до розгорнутої моделі. Загальний процес стиснення розширюється початковим процесом пошуку та навчання моделі, що може потребувати надзвичайно великих витрат ресурсів та часу. З'являються невизначені проблеми у процесі порівняння двох докорінно різних підходів та моделей. Питання сумісності та стандартизації залишаються невирішеними для навчальних кодеків, які розробляються незалежно один від одного з використанням різних архітектур, конвеєрів навчання та наборів даних, що ускладнює їхнє пряме порівняння та гальмує впровадження в галузі[8].

Історично навчене стиснення асоціюється лише з наскрізними моделями. Включення в огляд будь-якого методу стиснення (в якому для побудови, адаптації або керування кодерами, декодерами, перетвореннями, квантувальниками, ентропійними моделями або словниками використовуються машинне навчання, нейронні мережі, еволюційні обчислення, навчання з підкріпленням або будь-які інші механізми) дозволяє отримати розширену картину, що включає наступні моделі та методи:

- наскрізні навчені моделі із втратами для зображень, відео та аудіорекострукції;
- моделі навченого стиснення без втрат, що включає нейронні контекстні моделі, системи змішаного контексту та стиснення за допомогою великих мовних моделей у поєднанні з арифметичним кодуванням;
- навчені неявні нейронні представлення, що використовуються як формат стиснення та перенавчання на конкретних даних;
- генеративне та перцептивно-орієнтоване стиснення за допомогою генеративних мереж та дифузійних моделей;
- навчені компоненти, вбудовані в класичні кодеки — навчене внутрішньокадрове передбачення, фільтри, що діють у циклі обробки, регулювання швидкості передачі даних, навчені словники та рішення

кодера, що базуються на методах навчання з підкріпленням;

- метаевристичний підхід до проектування кодеків, в якому генетичні алгоритми, еволюційні стратегії та генетичне програмування забезпечують еволюцію кодів книг, вейвлет-фільтрів і навіть цілих програм, що генерують зображення;

- навчені міждоменні компресори для хмар точок, 3D-сцен, геномів, даних наукових моделювань, баз даних, навчених структур даних та самих ваг нейронних мереж.

Основна проблема порівняння вище наведених моделей полягає в надмірно розгалуженому застосуванні даних підходів та різному принципі роботи підходів. Зокрема, спеціалізована система по стисненню геномів є непорівнюваною із моделлю для загального стиснення. У даному разі порівняння слід проводити лише після поділу моделей та методів на окремі групи та виділенні найновіших та найбільш успішних представників даних груп.

### **Структурні компоненти та основоположні методи навченого стиснення**

Загальний принцип стиснення залишається таким же, як і в класичних методах – вилучення надлишковості, що присутнє в даних, приводячи стиснені представлення до межі ентропії. Саме еквівалентність між імовірнісним моделюванням та кодуванням перетворює будь-який навчений предиктор на компресор. Зокрема, модель, яка видає умовні ймовірності, подає дані до арифметичного кодера, коли очікувана довжина коду відповідає процесу стиснення. Отже, навчання нейронної мережі – це буквально навчання компресора з прямою метою дійти до ентропійної межі: функція втрат і довжина коду є однією й тією ж величиною. Саме тому прогрес у галузі навченого стиснення так тісно пов'язаний із прогресом у генеративному моделюванні: покращення оцінки щільності, авторегресійне моделювання, нормалізуючі потоки та варіаційне виведення, що призводить до знаходження рівнів, що наближаються до ентропійної межі  $H[X] = E[-\log_2 P(x)]$  [9].

Наскрізнi навчені кодеки будуються безпосередньо на зміщеній границі із врахуванням. У кодуванні з нелінійним перетворенням кодер відображає латентні величини, які квантуються та кодуються за ентропією відповідно до навченої моделі; метою навчання є функція Лагранжа, що описує залежність ступеню стиснення від спотворення  $E[-\log_2 P(z')] + \lambda E[\rho(x, g(z'))]$  [10]. У класичній моделі  $R(D) = \min I(X; X')$ , передбачається, що метою є низький рівень спотворення. Навчені генеративні кодеки показали, що це припущення не справджується для людських спостерігачів: мінімізація середньоквадратичної помилки (MSE) призводить до розмитих, нереалістичних реконструкцій [11]. Обмеження ліміту якості сприйняття занадто, як правило, призводить до підйому кривої спотворення до рівня стиснення, що, відповідно, вимагає компромісу або щодо рівня, або щодо спотворення. У загальному випадку межа ентропії є асимптотично досяжною і ніколи не перевищуваною. Навчання з даних визначає, наскільки близько практично можливо наблизитися до цієї межі, а розширення з втратами та введення перцепційного розширення визначають, де саме пролягає ця межа.

Основна проблема навченого стиснення лежить у межах знаходження та вибору моделей визначення перцептивно релевантної інформації та знаходження необхідного відношення рівня стиснення до спотворення відповідно до визначених обмежень. Для ефективного рішення даної задачі виведено конвеєр з нелінійного перетворення, квантування та ентропійного кодування, причому кожен компонент може бути замінений наскрізною системою або набором компонентів машинного навчання та нейронних мереж. Системи навченого стиснення розкладаються на кілька окремих структурних компонентів, кожен із яких виконує відповідну функцію: - перетворення аналізу (Analysis transform); - перетворення синтезу (Synthesis transform); - квантизація (Quantization); - ентропійна модель (Entropy model); - ентропійне кодування (Entropy coding).

Перетворення аналізу та синтезу відрізняються від традиційних лінійних пере-

творень, оскільки це нелінійні нейронні мережі, навчені на даних, що дозволяють здійснювати відображення у латентний простір, який краще піддається стисненню. Квантування призводить до появи недиференційованості, яку під час навчання усувають за допомогою неперервних наближень, таких як адитивний рівномірний шум, або через прямі моделі [12, 13, 14]. Основоположний елемент ентропійної моделі – це автокодувальник, нейронна мережа із вузьким місцем для знаходження латентного представлення, що може використовуватись як повноцінний елемент стиснення із втратами [15]. Але через значні обмеження автокодувальники залишилися лише орієнтовним архітектурним компонентом. Історично сама ентропійна модель була розвинута та модифікована низкою складових, які відрізняються тим, які кореляції вони враховують і як функціонують [13, 16].

Гіперприор: окремий гіперкодер/декодер передає додаткову інформацію для оцінки параметрів розподілу (наприклад, дисперсії), що змінюються в просторі для латентних змінних.

Авторегресійна контекстна модель: враховує локальні просторові залежності шляхом прогнозування кожної прихованої величини на основі причинно-наслідково декодованих сусідніх величин, що вимагає послідовної обробки.

Контекст за каналами: розділяє латентні вектори між каналами, використовуючи раніше декодовані фрагменти як контекст для наступних для врахування міжканальних взаємозв'язків.

Глобальна увага (трансформери): виявляє просторові залежності на великих відстанях, які пропускають локальні контекстні моделі, використовуючи механізми трансформера або уваги [17].

### **Моделі стиснення без втрат та моделі для загального стиснення**

Сучасне навчання безвтратного стиснення еволюціонувало від систем, що підтверджували концепцію, до різноманітної екосистеми, яка охоплює зображення, текст, аудіо, хмари точок, геноміку та наукові дані. Дедалі частіше як ключові елементи

використовуються великі мовні моделі із глибоким зв'язок між прогнозуванням та стисненням. Хоча коефіцієнти стиснення стабільно перевищують показники традиційних кодеків, обчислювальна ефективність та узагальнення для різних бенчмарків залишаються головними перешкодами для широкого впровадження цієї технології. Огляди та порівняльні тести почали систематизувати оцінку, але все ще існує потреба у всебічних міжмодальних порівняльних тестах для навчання стиснення без втрат [18, 19, 20, 21].

Дослідження та тести порівняльної оцінки у публікації, представлені у корпусі NNLCB [18], дають цінну, але неповну картину. Корпус оцінює 17 компресорів на 28 наборах даних за допомогою 19 уніфікованих показників, пропонуючи всебічну систему оцінки, але не враховує нових підходів на основі великих мовних моделей. У роботі оцінено основні моделі NNCP [22], PAC [23], TRACE [24] та Dzip [25], які на узагальненому корпусі даних в середньому отримали від 8% до 42% приросту ефективності стиснення в порівнянні із класичними моделями, але разом з тим жертвуючи швидкістю стиснення та відновлення. Аналіз NNLCB межує з появою нового підходу загального стиснення на базі великих мовних моделей LMCompress [26], що в середньому переважає існуючі методи станом на 2025 рік та отримує на зображеннях, аудіо, відео та текстових даних коефіцієнт стиснення в 2-4 рази більший за наявні моделі, що слугує основною мотивацією порушення питання досяжності ентропійної межі Шеннона. Автори стверджують, що мовні моделі спираються на здатність до апроксимізації необчислюваної індукції Соломонова, яка відкриває нову парадигму стиснення за Колмогоровим, що пояснюється здатністю моделей розуміти дані, що стискаються [26]. Дане твердження базується лише на емпіричних результатах, тому не може бути спростованим чи підтвердженим, але порушує нову проблему знаходження підходу оцінки ентропійної межі для нових систем.

Результати використання мовних моделей були значно покращені в моделі Nacrih [27] з попередньо навченою мо-

деллю з ансамблевим змішуванням контексту, що забезпечують високий ступінь стиснення, але за рахунок спеціалізації на текстових даних. Проведено аналіз меж ентропії Шеннона у порівнянні з фактичними розмірами стиснених файлів `alice29.txt` (152 089 байт). `Nacrith` стискає файл на 80 % менше за межу 0-го порядку, на 73 % менше за межу 1-го порядку, і на 63 % менше за межу 2-го порядку, досягаючи коефіцієнта стиснення 8.71 та 8.92 на датасеті `enwik8`. Але дані результати не можна повноцінно порівнювати із межами ентропії через відсутність врахування потреби завантаження ваг моделі на сторону отримувача, що перевищує 500Мб та є більшою за стискувальні набори даних.

Остання модель `OmniZip`[27] претендує на кросдоменність та універсальність, бо отримує високі показники стиснення порівняно з іншими спеціалізованими моделями. Модель перевірено на даних зображень, тексту, мови, медичних, тактильних, генетичних та узагальнених баз даних. Модель переважає класичні методи стиснення загального типу від 10% до 70% за різними модальностями, але не досягає найкращих коефіцієнтів стиснення в жодній. Система використовує полегшену магістральну мережу `RWKV-7`, разом із рядом оптимізаційних підходів, що забезпечує оптимальне співвідношення ефективності стиснення та швидкості обробки даних. Основним обмеженням прискорювача `LLM` є те, що він призначений виключно для офлайн - інференції та не підтримує вбудованої обробки онлайн-даних або потокового аудіовізуального вхідного сигналу.

### Моделі стиснення із втратами та особливі випадки

Галузь навченого стиснення з втратами охоплює низку архітектурних категорій – від автоенкодерів і варіаційних моделей до генеративно-змагальних мереж, дифузійних моделей та неявних нейронних представлень, кожна з яких має власні огляди та тести, що оцінюють найсучасніші методи[28]. За період з 2010 до 2024 року

кількість публікацій про нейронне стиснення зросло до понад 100 публікацій за рік, враховуючи, що станом на 2023 рік 2000 публікацій є доменно орієнтованими. Стиснення з втратами є достеменно спеціалізованими на певний домен для максимізації ефективності моделей та методів, що попередньо унеможлиблює класифікацію та порівняння. Перший крок до симпліфікації порівняння – розділ даних на медіа дані та спеціалізовані наукові дані.

Комплексний аналіз підходів до стиснення із втратами наукових збірок даних [29] оцінює застосування понад 46 змішаних методів стиснення з втратами загального призначення, а виділяє окремі випадки для кожного окремого виду галузевих даних, що свідчить про фрагментарну систематизацію: компресори розробляються як окремі моделі, оцінюються на тестах у вузьких областях застосування та рідко перевіряються на міждоменну генералізацію. Надмірна спеціалізація за сферами застосування є дуже поширеним явищем: більшість навчальних моделей стиснення розроблені під конкретні типи моделювання (наприклад, космологія, турбулентність, клімат) і не можуть бути застосовані до нових розподілів даних.

Медіадані – зображення, відео, аудіо із застосуванням сучасних показників сприйняття – крім класичних `PSNR` та `MS-SSIM` – дозволило досягти значного підвищення ефективності стиснення завдяки збільшенню допустимого рівня непомітних спотворень при нижчих бітрейтах.

Спираючись на результати нещодавніх оглядових досліджень у галузі навчання методів стиснення зображень та відео[30], отримано сучасні кодеки із парето-збалансованими параметрами, наведеними у таблиці 1.

Нейронне стиснення із втратами досягнуло рівня отримання стандартизованого `ISO JPEG AI`[31], середнє зниження швидкості передачі даних перевищує 30 % порівняно з інтракодуванням `VVC`, що демонструє рух від академічних прототипів до рівня міжнародного стандарту.

Таблиця 1.

Порівняльна таблиця моделей

| Назва моделі | Дані       | Датасет                         | Опис переваги                                                                                                                                                      |
|--------------|------------|---------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| SAAF         | зображення | Kodak, CLIC, Tecnick            | Найкращий вибір для якості, -20,57% BD коефіцієнту порівняно з VTM-9.1, 67 мс затримки; 123М параметрів                                                            |
| PICO         | зображення | Kodak, CLIC 2020, DIV2K         | Найкращі результати на мобільних девайсах; x3 покращення стиснення за традиційні кодеки без змін якості; 150-230 мс затримки; затрати пам'яті не перевищують 40 Мб |
| LALIC        | зображення | Kodak, CLIC, Tecnick.           | Найкраща модель з меншою точністю; -17.63% BD коефіцієнту порівняно з VTM-9.1;                                                                                     |
| DCVC-UF      | відео      | UVG, MCL-JCV, HEVC Classes B-E  | Найкращі показники пропускної здатності; -42,2% середнього BD коефіцієнту порівняно з VTM-17; затримка 370 мс.                                                     |
| UPC/UIC      | відео      | UVG, MCL-JCV, HEVC Classes B-E  | Збалансована модель -12.1% BD коефіцієнту порівняно з DCVC-RT; 46 М параметрів, 65/46 fps                                                                          |
| DCVC-RT      | відео      | UVG, MCL-JCV, HEVC Classes B-E. | Найкраща практична модель, Reported 21% виграш стиснення порівняно з H.266/VTM та 125/113 fps                                                                      |

Висновки

За відносно короткий період розвитку в 10 років у галузі навченого стиснення переважають результуючі три парадигми: використання мовних моделей для загального стиснення без втрат, що потенційно, але не підтверджено долають ентропійні межі; використання метрик та характеристик визначення перцептивно релевантних елементів у аудіо, відео та звукових даних для підбору кращої модифікованої кривої стиснення-втрати-сприйняття; використання вузькоспрямованих методів для досягнення надзвичайних непорівнюваних результатів у спеціалізованих випадках.

Використання великих мовних моделей потребує чіткішої моделі визначення ентропії та загального перегляду процесу стиснення, що комбінує нові можливості мовних моделей, які мають характер розуміння даних, у одну зв'язну ентропійну модель. Загальної моделі стиснення даних досі не знайдено, що лишає місце для подальшого дослідження оптимізації моделей та підвищення мультимодальності. Моделі стиснення із втратами потребують чіткі-

шого визначення параметрів перцептивно важливих елементів для підвищення можливих досяжних меж стиснення, проблема оптимізації процесів навченого стиснення лишається відкритою. Спеціалізовані набори даних потребують механізму уніфікації та чіткого визначення меж перцептивності та релевантності елементів даних, хоча невизначеність рамок надає більше простору для розвитку та тестування нових методів, що продукують абсурдно високі параметри.

Передові моделі загального стиснення (OmniZip), сертифіковані моделі на базі машинного навчання (JPEG AI) є прикладами проміжного етапу розвитку навченого стиснення та ефективного застосування методів та моделей. Але питання появи докорінно нових підходів до стиснення із комбінованими та новими інтелектуальними підходами залишається відкритим.

References

1. LI, Ming, et al. Understanding is compression. 2024.WELCH, Terry A. A technique for high-performance data compression. *Computer*, 1984, 17.06: 8-19.

2. WILKENFELD, Daniel A. Understanding as compression: DA Wilkenfeld. *Philosophical Studies*, 2019, 176.10: 2807-2831
3. YANG, Yibo; MANDT, Stephan; THEIS, Lucas. An introduction to neural data compression. *Foundations and Trends in Computer Graphics and Vision*, 2023, 15.2: 113-200
4. Huang, C.-H. and Wu, J.-L. (2024). Unveiling the Future of Human and Machine Coding: A Survey of End-to-End Learned Image Compression. *Entropy*, 26. doi: [10.3390/e26050357](https://doi.org/10.3390/e26050357).
5. Hu, Y., Yang, W., Z. and Liu, J. (2020). Learning End-to-End Lossy Image Compression: A Benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44, pp. 4194-4211. doi: [10.1109/tpami.2021.3065339](https://doi.org/10.1109/tpami.2021.3065339).
6. Cao, M., Dai, W., Li, S., Li, C., Zou, J., Chen, Y. and Xiong, H. (2025). End-to-End Optimized Image Compression With Deep Gaussian Process Regression. *IEEE Transactions on Circuits and Systems for Video Technology*, 35, pp. 3770-3785. doi: [10.1109/tcsvt.2022.3209661](https://doi.org/10.1109/tcsvt.2022.3209661).
7. Ehrlich, M. (2022). The First Principles of Deep Learning and Compression. *ArXiv*, abs/2204.01782. doi: [10.48550/arxiv.2204.01782](https://doi.org/10.48550/arxiv.2204.01782).
8. Porat, D. and Levi, D. (2026). Learned Image and Video Compression: Foundations, Algorithms, and Open Challenges. *Proceedings of the 5th Mile-High Video Conference*. doi: [10.1145/3789239.3793287](https://doi.org/10.1145/3789239.3793287).
9. Del'etang, G., Ruoss, A., Duquenne, P.-A., Catt, E., Genewein, T., Mattern, C., Grau-Moya, J., Li, W., Aitchison, M., Orseau, L., Hutter, M. and Veness, J. (2023). Language Modeling Is Compression. *ArXiv*, abs/2309.10668. doi: [10.48550/arxiv.2309.10668](https://doi.org/10.48550/arxiv.2309.10668).
10. Agustsson, E. and Theis, L. (2020). Universally Quantized Neural Compression. *ArXiv*, abs/2006.09952. doi: [10.48550/arxiv.2006.09952](https://doi.org/10.48550/arxiv.2006.09952).
11. Blau, Y. and Michaeli, T. (2019). Rethinking Lossy Compression: The Rate-Distortion-Perception Tradeoff., pp. 675-685. doi: [10.48550/arxiv.1901.07821](https://doi.org/10.48550/arxiv.1901.07821).
12. Ballé, J., Laparra, V. and Simoncelli, E. P. (2016). End-to-end Optimized Image Compression. *ArXiv*, abs/1611.01704. doi: [10.48550/arxiv.1611.01704](https://doi.org/10.48550/arxiv.1611.01704).
13. Wang, C., Liang, Z., Niu, K. and Zhang, P. (2026). A Theoretical Framework for Rate-Distortion Limits in Learned Image Compression. *ArXiv*, abs/2601.09254. doi: [10.48550/arxiv.2601.09254](https://doi.org/10.48550/arxiv.2601.09254).
14. Sheng, X., Li, L., Liu, D. and Li, H. (2024). Spatial Decomposition and Temporal Fusion Based Inter Prediction for Learned Video Compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 34, pp. 6460-6473. doi: [10.1109/tcsvt.2024.3360248](https://doi.org/10.1109/tcsvt.2024.3360248).
15. Lesyk, V.O. and Doroshenko, A.Yu. (2023) 'Image compression module based neural network autoencoders', *Problems in Programming*, 1, pp. 48-57. [in Ukrainian]
16. Chen, T., Z., Shen, Q., Cao, X. and Wang, Y. (2019). End-to-End Learnt Image Compression via Non-Local Attention Optimization and Improved Context Modeling. *IEEE Transactions on Image Processing*, 30, pp. 3179-3191. doi: [10.1109/tip.2021.3058615](https://doi.org/10.1109/tip.2021.3058615).
17. Jiang, W., Yang, J., Zhai, Y., Gao, F. and Wang, R. (2023). MLIC++: Linear Complexity Multi-Reference Entropy Modeling for Learned Image Compression. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21, pp. 1 - 25. doi: [10.1145/3719011](https://doi.org/10.1145/3719011).
18. Sun, H., H., Ling, F., Xie, H., Sun, Y., Yi, L., Yan, M., Zhong, C., Liu, X. and Wang, G. (2025). A survey and benchmark evaluation for neural-network-based lossless universal compressors toward multi-source data. *Frontiers of Computer Science*, 19. doi: [10.1007/s11704-024-40300-5](https://doi.org/10.1007/s11704-024-40300-5).
19. Hishida, K., Liu, C., Paparrizos, J. and Elmore, A. (2025). Beyond Compression: A Comprehensive Evaluation of Lossless Floating-Point Compression. *Proc. VLDB Endow.*, 18, pp. 4396-4409. doi: [10.14778/3749646.3749701](https://doi.org/10.14778/3749646.3749701).
20. Li, Y., Guo, Y., Guerin, F. and Lin, C. (2024). Evaluating Large Language Models for Generalization and Robustness via Data Compression. *ArXiv*, abs/2402.00861. doi: [10.48550/arxiv.2402.00861](https://doi.org/10.48550/arxiv.2402.00861).
21. Zhang, J., Cheng, Z., Zhao, Y., Wang, S., Zhou, D., Lu, G. and Song, L. (2024). L3TC: Leveraging RWKV for Learned Lossless Low-Complexity Text Compression., pp. 13251-13259. doi: [10.48550/arxiv.2412.16642](https://doi.org/10.48550/arxiv.2412.16642).
22. Bellard, F. (2021). NNCP v2: Lossless Data Compression with Transformer..
23. Mao, Y., Li, J., Cui, Y. and Xue, J. (2023). Faster and Stronger Lossless Compression with Optimized Autoregressive

- Framework. 2023 *60th ACM/IEEE Design Automation Conference (DAC)*, pp. 1-6. doi: 10.1109/dac56929.2023.10247866.
24. Mao, Y., Cui, Y., Kuo, T.-W. and Xue, C. (2022). TRACE: A Fast Transformer-based General-Purpose Lossless Compressor. *Proceedings of the ACM Web Conference 2022*. doi: 10.1145/3485447.3511987.
  25. Goyal, M., Tatwawadi, K., Chandak, S. and Ochoa, I. (2020). DZip: Improved General-Purpose Lossless Compression Based on Novel Neural Network Modeling. *2020 Data Compression Conference (DCC)*, pp. 372-372. doi: 10.1109/dcc47342.2020.00065.
  26. Li, Z., Huang, C., Wang, X., Hu, H., Wyeth, C., Bu, D., Yu, Q., Gao, W., Liu, X. and Li, M. (2024). Lossless data compression by large models. *Nature Machine Intelligence*, 7, pp. 794 - 799. doi: 10.1038/s42256-025-01033-7.
  27. Tacconelli, R. (2026). Nacrih: Neural Lossless Compression via Ensemble Context Modeling and High-Precision CDF Coding. *ArXiv*, abs/2602.19626. doi: 10.48550/arxiv.2602.19626.
  28. Jamil, S., Piran, M. and Rahman, M. (2022). Learning-Driven Lossy Image Compression; A Comprehensive Survey. *ArXiv*, abs/2201.09240. doi:10.48550/arxiv.2201.09240.
  29. Di, S., Liu, J., Zhao, K., Liang, X., Underwood, R., Zhang, Z., Shah, M., Huang, Y., Huang, J., Yu, X., Ren, C., Guo, H., Wilkins, G., Tao, D., Tian, J., Jin, S., Jian, Z., Wang, D., Rahman, M. H., Zhang, B., Calhoun, J. C., Li, G., Yoshii, K., Alharthi, K. and Cappello, F. (2024). A Survey on Error-Bounded Lossy Compression for Scientific Datasets. *ACM Computing Surveys*, 57, pp. 1 - 38. doi: 10.1145/3733104.
  30. Wondmagegn, A. B., Dinh, T. T. T., Win, T. T., Won, D. and Cho, S. (2026). A Survey on Deep Learning-Based Image Compression: Methods, Efficiency, and Open Challenges. *2026 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC)*, pp. 108-111. doi: 10.1109/icaic68212.2026.11454276.
  31. Pinheiro, A. M. G. (2025). JPEG Column: 106th JPEG Meeting. *ACM SIGMultimedia Records*, 17, pp. 1 - 1. doi: 10.1145/3810993.3810996.
- Дата першого надходження до видання: 14.07.2026  
 Внутрішня рецензія отримана: 03.08.2026  
 Зовнішня рецензія отримана: 05.08.2026  
 Дата прийняття до друку: 10.08.2026  
 Дата публікації: 29.08.2026
- Про авторів:**
- <sup>1</sup>Лесик Валентин Олександрович,  
аспірант, асистент кафедри  
*Lesyk Valentyn*,  
PhD student  
<https://orcid.org/0000-0002-8307-5707>.
- <sup>1,2</sup>Дорошенко Анатолій Юхимович,  
доктор фізико-математичних наук,  
професор  
*Doroshenko Anatoliy*,  
Ph.D. ( physics and mathematics),  
professor  
<http://orcid.org/0000-0002-8435-1451>.
- Місце роботи авторів:**
- <sup>1</sup>Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»  
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
- <sup>2</sup>Інститут програмних систем  
НАН України  
Institute of Software Systems  
NAS of Ukraine