Vol. 7, No. 1, 2025

Олена Станкевич[1], Данило Матвійків[2]

[1] Кафедра систем автоматизованого проектування, Національний університет "Львівська політехніка", вул. С. Бандери, Львів, Україна, E-mail: olena.m.stankevych@lpnu.ua, ORCID 0000-0002-5977-6351
[2] Кафедра систем автоматизованого проектування, Національний університет "Львівська політехніка", вул. С. Бандери, Львів, Україна, E-mail: danylo.o.matviikiv@lpnu.ua, ORCID 0009-0009-2643-2809

САМОНАВЧАЛЬНІ ТРАНСФОРМЕРИ ЗОРУ ДЛЯ КРОС-МОДАЛЬНОГО НАВЧАННЯ (ОГЛЯД)

Отримано: Лютий 28, 2025 / Переглянуто: Березень 28, 2025 / Прийнято: Квітень 01, 2025

https://doi.org/

Анотація. Системи комп'ютерного зору все більше розширюють межі свого застосування у задачах аналізу візуальних даних. Найбільшого розвитку та вдосконалення зазнають методики навчання моделей, адже результати цього етапу суттєво впливають на кінцеву класифікацію об'єктів та інтерпретацію вхідної інформації.

Зазвичай у системах комп'ютерного зору використовують для навчання згорткові нейронні мережі (Convolution Neural Network, CNN). Недоліками таких систем є суттєві обмеження у крос-модальному навчанні, реалізації мультимодальності, маркуванні даних значних обсягів тощо. Одним зі шляхів подолання цих перешкод є використання трасформерів зору (Vision Transformers, ViT), які порівняно з класичними CNN мають вищу продуктивність, завдяки зниженим індуктивним упередженням та високій ефективності паралельних обчислень. Впровадження технологій самонавчання (Self-Supervised Learning, SSL) дає змогу суттєво зменшити залежність від вручну маркованих даних, сприяючи формуванню узагальнених представлень про зображення. Крос-модальне навчання (Cross-Modal Learning, CML) розширює можливості їх опрацювання, об'єднуючи дані різних типів. Розроблення нового підходу, який би поєднував можливості крос-модального навчання та самонавчання ViT у єдиній архітектурі, забезпечить адаптивність, ефективність та масштабованість системи загалом у різних сферах застосування.

Метою виконаних досліджень є детальний огляд ViT, підходів до їх архітектури та методів підвищення їх ефективності. Розглянуто математичні основи основних концепцій ViT, крос-модального навчання та самонавчання, відомі модифікації ViT та їх інтеграцію з технологіями SSL і CML. Наведено порівняння методів за характеристиками, продуктивністю та ефективністю. Окреслено ключові виклики, які стоять перед дослідниками та розробниками, та перспективи створення універсальних моделей у галузі комп'ютерного зору.

ViT змінюють комп'ютерний зір, фіксуючи глобальні залежності в зображеннях. Незважаючи на певні проблеми, ViT забезпечують чудову масштабованість і продуктивність для великих наборів даних. Активний пошук методів подолання існуючих обмежень сприяє тому, що ViT можуть стати головним інструментом для вдосконалення класифікації зображень, виявлення об'єктів та інших завдань комп'ютерного зору.

Ключові слова: трансформери зору; самонавчання; крос-модальне навчання; комп'ютерний зір; глибоке навчання

Вступ

Комп'ютерний зір спрямований на те, щоб дати можливість комп'ютерам осмислено інтерпре-тувати візуальну інформацію з навколишнього світу, імітуючи можливості людського зору

без втрати точності та з більшою ефективністю [1]. Прогрес комп'ютерного зору впродовж останнього десяти-ліття зумовлений трьома чинниками: (*1*) розвитком глибокого навчання (Deep Learning, DL), яке забезпечує наскрізне вивчення дуже складних функцій на основі необроблених даних; (*2*) досягненням локалізованих обчислювальних потужностей за допомогою графічних процесорів та (*3*) відкритим доступом до великих маркованих наборів даних, на яких можна тренувати ці алгоритми. Водночас сучасні системи комп'ютерного зору, попри значні досягнення, мають суттєві обмеження у крос-модальному навчанні.

Згорткові нейронні мережі (Convolution Neural Network, CNN), які традиційно використовують у задачах комп'ютерного зору, мають високу ефективність для вилучення локальних ознак, однак їхня здатність до глобального контекстного аналізу обмежена через архітектурні особливості. Це робить CNN малопридатними для узгодження різних типів даних, наприклад, текстових описів із візуальними даними. Для подолання цього недоліку часто застосовують складні методи постобробки або спеціалізовані моделі, що суттєво ускладнює інтеграцію в реальних умовах.

Підхід до обробки даних у багатьох застосуваннях змінили трансформери зору (Vision Transfor-mers, ViT), які спочатку використовували в задачах обробки природної мови (Nature Language Processing, NLP) для машинного перекладу, генерації та резюмування тексту, аналізу настроїв тощо. Згодом їх почали активно впроваджувати в системи комп'ютерного зору для розпізнавання зображень [2].

ViT, завдяки наявності механізму самоуваги (*Self-Attention*), можуть моделювати довгострокові залежності між об'єктами на зображеннях, що підвищує їх ефективність порівняно з CNN. Однак більшість ViT працюють в одномодальному режимі, тобто аналізують лише візуальні дані без інтеграції з іншими модальностями, такими, як текст, аудіо чи сенсорні дані. Така ізольованість є суттєвим недоліком у застосуваннях, які потребують комплексного опрацювання інформації, наприклад, в автономних системах чи асистентах штучного інтелекту для одночасного розпізнавання зображень та обробки текстових інструкцій. На сьогодні перспективним напрямком є використання ViT у мультимодальних моделях (співставлення тексту, зображень та інших типів даних) для інтеграції різних джерел інформації. Однак і тут є певні труднощі, які вимагають розроблення підходів до їх вирішення.

Ще одним викликом є проблема маркування даних, яка є особливо критичною в мультимодальних системах. Зокрема, методи навчання з учителем вимагають наявності великих обсягів маркованих даних, що є обмеженням для багатьох задач. Наприклад, узгодження медичних зображень із відповідними діагностичними звітами або прив'язка відеофрагментів до текстових описів є трудомісткими завданнями, які потребують ручної роботи експертів. Це значно сповільнює розвиток та впровадження мультимодальних систем штучного інтелекту в реальних застосуваннях.

Окрім проблеми маркування, сучасні мультимодальні моделі часто розділяють обробку різних модальностей, що призводить до складних і малоефективних процесів інтеграції. Наприклад, великі мовні моделі (Large Language Models, LLMs) можуть досягати високої продуктивності у текстових завданнях, але без візуального контексту вони не здатні повною мірою аналізувати реальні зобаження. Натомість ViT легко справляються з аналізом зображень, але без текстової семантики їхня здатність до глибшого розуміння контенту залишається обмеженою.

Окрему проблему становить неефективність традиційних архітектур у мультимодальному навчанні. CNN, через свою схильність до локального аналізу, не здатні забезпечити ефективне узгод-ження між модальностями, тоді як ViT, працюючи в ізольованому середовищі, не використовують потенціал комбінованого навчання. Наприклад, під час аналізу фотографій природних катастроф системи, які працюють лише з візуальними даними, не здатні корелювати їх із текстовими повідом-леннями або іншими даними, що обмежує можливості аналізу ситуації в режимі реального часу.

Отже, існуючі підходи мають низку критичних обмежень:

- CNN неефективні у крос-модальному навчанні через локальність обробки ознак:

- ViT у стандартному вигляді працюють тільки з одним типом даних, не залучаючи потенціал мультимодальних зв'язків:
- мультимодальні системи вимагають великих обсягів маркованих даних, що робить їх навчання дорогим та непрактичним.
- існуючі методи поєднання модальностей досить складні, малоефективні та обмежують створення гнучких систем штучного інтелекту.

Водночас набувають популярності методи самонавчання (Self-Supervised Learning, SSL), які дають змогу моделі вивчати узагальнені уявлення про дані на немаркованих вибірках [3]. Такі попередньо навчені моделі можна згодом донавчати для розв'язання конкретних завдань, заощаджуючи ресурси та час на маркування. Ще один перспективний напрям – крос-модальне навчання (Cross-Modal Learning, CML), яке поєднує дані з різних джерел (наприклад, візуальні та текстові), що значно розширює застосування моделей у реальних завданнях [4].

Отже, вирішення зазначених проблем вимагає нового підходу, який би поєднував можливості крос-модального навчання та самонавчання ViT в єдиній архітектурі, що забезпечить адаптивність, ефективність та масштабованість у різних сферах застосування.

Постановка проблеми

Метою виконаних досліджень є всебічний огляд ViT, який охоплює існуючі підходи до їх архітектури, математичні основи ключових концепцій ViT, крос-модального навчання та самонавчання, основні модифікації ViT та їх інтеграцію з технологіями SSL і CML. Порівняння методів за характеристиками, продуктивністю та ефективністю дають глибоке розуміння стану проблеми та визначення перспективних напрямків її розв'язання.

Виклад основного матеріалу

Архітектурні принципи ViT. Vision Transformers (ViT) – це технологія адаптації архітектури текстового трансформера для аналізу зображень. Замість піксельних матриць модель оперує послідовністю сегментів зображення, подібно до послідовності слів у текстовому трансформері. Вхідне зображення розбивають на невеликі сегменти фіксованого розміру, наприклад 16×16 пікселів [2]. Кожний сегмент розгортається у вектор (шляхом конкатенації значень пікселів) та лінійно відображається у послідовність блоків (*Embeddings*) певної розмірності. Запишемо зображення X у вигляді

$$X \in R^{H \times W \times C}, \tag{1}$$

де $H \times W$ – розмір, C – кількість каналів. Якщо сегмент має розмір $P \times P$, то їх кількість можна обчислити за формулою

$$N = \frac{H \times W}{P^2}. \tag{2}$$

Кожний сегмент x_i (розмірність $P^2 \times C$) відображається лінійним шаром у вектор $z_i = W \cdot flatten(x_i) + b$, де W – матриця ваг, а b – зміщення. Таким чином формується послідовність векторів $\{z_i\}_{i=1}^{N}$, до якої додають позиційні ознаки (*Learnable Position Embeddings*), щоб модель враховувала положення сегмента в зображенні.

До цієї послідовності звичайно додають спеціальний токен класу [*CLS*] – вектор, призначений для агрегування глобальної інформації, необхідної для прогнозування класу зображення. Отримана послідовність $[CLS, z_1, z_2, ..., z_N]$ подається на стандартний трансформер-кодувальник (*Encoder*), який складається з шарів мультиголової самоуваги (*Multi-Head Self-Attention*) і позиційних шарів прямого зв'язку (*FeedForward Networks*). Основним механізмом ViT є самоувага (*Self-Attention*), яка

дає змогу кожному сегменту взаємодіяти з іншими, враховуючи глобальний контекст зображення. Математично самоувагу записують у вигляді

$$Attention(Q,K,V)=soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{3}$$

де Q, K, V – матриці запитів (*Queries*), ключів (*Keys*) та значень (*Values*), отримані перетворенням вихідних блоків, відповідно, а d_k – розмірність головних векторів [5].

За допомогою рівняння (3) можна обчислити вагові коефіцієнти між усіма парами сегментів: спочатку скалярні добутки QK^T оцінюють суміжність (подібність) між кожним запитом і ключем, потім ділення на $\sqrt{d_k}$ стабілізує масштаби градієнтів, а функція *softmax* нормалізує ваги. Отримані ваги застосовують до значень V, таким чином агрегуючи інформацію з усіх сегментів. Мультиголова увага повторює цю процедуру паралельно для кількох наборів Q_i, K_i, V_i (*голів уваги*) і з'єднує результати, збагачуючи модель здатністю фокусуватися на різних аспектах образу одночасно.

Після кількох шарів самоуваги та нелінійних перетворень трансформер-кодувальника, вихідний вектор [*CLS*] використовують для класифікації зображення через повнозв'язний шар. Така трансформерна архітектура для зображень (рис. 1) без жодних згорткових шарів показала високу ефективність за наявності достатньо великого обсягу даних для попереднього навчання [2].

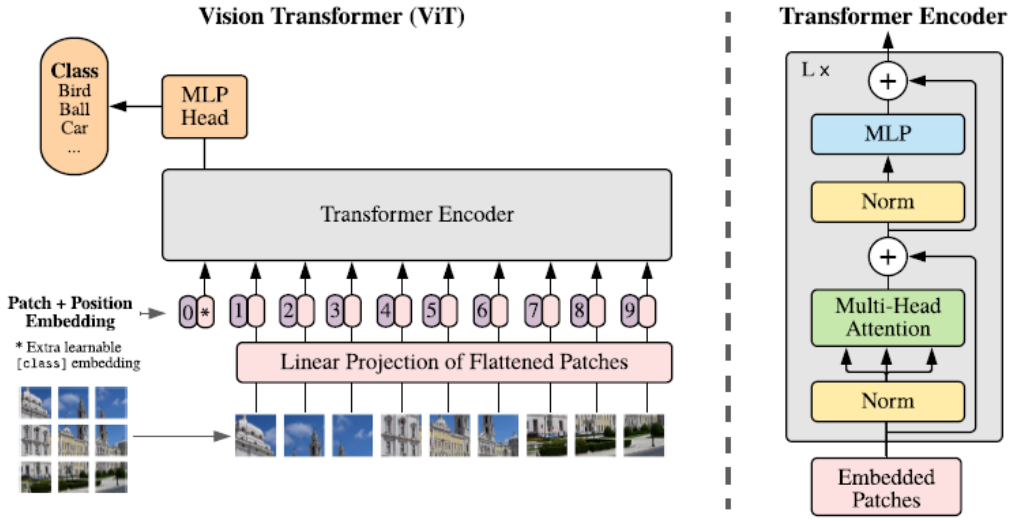

Рис. 1. Архітектура ViT.

Дослідження на наборах даних ImageNet та інших продемонстрували, що ViT перевершує найкращі CNN, водночас потребуючи менших обчислювальних ресурсів для навчання. Досягти цього вдалось завдяки навчанню на великій приватній колекції зображень (наприклад, JFT-300M) з наступним донавчанням (*Transfer Learning*) на цільових наборах (ImageNet, CIFAR-100 та ін.). Важливо, що ViT фактично усунув необхідність у згорткових шарах для побудови ефективних моделей комп'ютерного зору.

Отже, реалізація ViT знаменує трансформаційний перехід від традиційних згорткових моделей до підходу, який безпосередньо адаптує архітектуру трансформера, розробленого для обробки природної мови, до сфери комп'ютерного зору. Така методика переосмислює обробку зображень, розглядаючи зображення як послідовність сегментів, а не як суцільну матрицю пікселів. Завдяки цьому поділу модель може використовувати механізми самоуваги, отримуючи цілісне уявлення про зображення за один крок.

Порівняння основних моделей ViT. Після появи базової моделі ViT [6] запропоновано низку поліпшень та варіацій архітектури, спрямованих на підвищення ефективності та зменшення вимог до даних. Розглянемо декілька знакових моделей та їхні характеристики, які подано у Таблиці 1.

Таблиця 1

Порівняння моделей Vision Transformers та їхні характеристик

Модель (рік)	**Кількість параметрів**	**Дані для навчання**	**ImageNet Top-1**	**Ключові особливості**
ViT-B/16 (2021) [2]	~86 млн.	JFT-300M (попередньо) + ImageNet	~84% (налаштування)	Сегменти 16×16, глобальна увага, потребує великих даних
DeiT-B/16 (2021)[6]	~86 млн.	ImageNet-1K (тільки)	83,1% (без дистил.) 85.2% (з дистил.)	Ефективне навчання «з нуля», сегмент-дистилятор для знань CNN
Swin-B (2021) [7]	~88 млн.	ImageNet-1K	87,3%	Ієрархічні рівні, локальні «вікна» уваги, масштабованість на детектування/сегментацію

Трансформери зору *Data-Efficient Image Transformers (DeiT)* запропонували для ефективнішого навчання трансформерів «з нуля» на відносно невеликих наборах даних [6]. Автори показали, що ViT-архітектуру з ~86 млн параметрів можна навчити лише на ImageNet (1,2 млн зображень) протягом кількох днів і досягти достатньо високої точності 83,1% без зовнішніх даних. Для цього автори використали методику відокремлення (*Distillation*): введено спеціальний токен відокремлення (*Distillation Token*), який навчається відтворювати відповіді «вчителя» (попередньо навченої CNN). Завдяки цьому «учень»-трансформер переймає знання у згорткової моделі, покращуючи власну узагальнюючу здатність. Точність моделі на ImageNet вдалося підвищити до 85,2%, що можна співставити з найкращими CNN, але уникнувши попереднього навчання на зовнішніх базах. Отже, DeiT довів, що навіть без гігантських наборів даних ViT можуть успішно конкурувати з CNN, якщо використати оптимальні стратегії навчання, наприклад, відокремлення через самоувагу.

Окреме місце в еволюції ViT займає модель *Swin Transformer* [7]. Її особливістю є введення ієрархічної архітектури трансформера з поступовою агрегацією сегментів та обмеженням області самоуваги в межах локальних вікон (*Shifted Windows*). На перших шарах Swin Transformer розбиває зображення на малі вікна і обчислює самоувагу лише між сегментами всередині кожного вікна, що різко скорочує обчислювальну складність (до лінійної залежності від розміру зображення) та вводить корисний індуктивний додаток локальності, притаманний CNN. Під час переходу на глибші шари вікна зсуваються і зливаються, забезпечуючи зв'язність усієї картинки. Цей підхід дав можливість побудувати універсальну основу для різних задач комп'ютерного зору. Swin Transformer досяг 87,3% точності на наборі ImageNet-1K, перевершивши попередні SOTA-моделі приблизно на +2-3% абсолютної точності. Важливо, що Swin Transformer прекрасно масштабувався на задачі детекту-вання та сегментації, його застосування сприяло підвищенню точності на COCO (58,7% APbbox, 51,1% APmask) та ADE20K (53,5% mIoU). Це підтвердило потенціал ViT як універсальних архітектур для різних типів задач – від класифікації до прогнозування, де раніше домінували CNN.

Відомо також інші варіанти трансформерних архітектур. Наприклад, TNT (Transformer in Transformer) [8], PVT (Pyramid Vision Transformer) [9], CvT (Convolutional ViT) [10] – кожний із них

має свої вдосконалення: від використання вкладених трансформерів для опрацювання локальних деталей до поєднання згорток з механізмом самоуваги. Такі модифікації спрямовані на підвищення ефективності навчання на обмежених наборах даних, зменшення вимог до пам'яті та прискорення обчислень, зберігаючи водночас високу точність. Наукові експерименти свідчать, що належним чином спроєктовані трансформери можуть переважати CNN за якістю розпізнавання, особливо за умови доступності великих обсягів даних чи потужних механізмів самонавчання [2].

За даними у Таблиці 1, архітектурні удосконалення та нові навчальні стратегії сприяли тому, що ViT досягнули та навіть перевершили результати CNN на ImageNet. Важливо, що навіть менші за параметрами моделі можна навчити ефективно, якщо використовувати самонавчання або відокремлення.

Особливості моделей ViT (2023–2025). Упродовж 2020–2022 рр. закладено теоретичний фундамент ViT. Зокрема, подано перші успішні застосування трансформерів для вирішення завдань обробки природної мови [4], а також результати їх адаптації для комп'ютерного зору [2, 5–10]. У 2023–2025 рр. розвиток ViT суттєво прискорився через збільшення розмірів моделей (до десятків мільярдів параметрів), створення універсальних (*Foundation*) моделей, упровадження самонавчання та вдосконалення ефективності обчислень. У Таблиці 2 подано характеристики найвідоміших моделей, кожна з яких використовує різні методи для досягнення найвищих критеріїв продуктивності та ефективності. Проаналізуємо основні досягнення цього періоду.

Масштабування ViT до мільярдів параметрів. У праці [11] дослідники Google представили модель найбільшого на той час візуального трансформера ViT-22B із 22 млрд. параметрів. Ця модель у 5,5 рази більша, ніж попередня (ViT-e, ~4 млрд параметрів), і її вдалося успішно навчити завдяки покращенням у стабільності (зокрема, QK-нормалізація) та ефективності (асинхронні паралельні обчислення) у шарі самоуваги. Використання нової моделі дало змогу суттєво поліпшити продуктивність у багатьох задачах комп'ютерного зору (класифікація, детектування, сегментація та ін.) під час перенесення навчених ознак (*Transfer Learning*). Також, масштабування до 22B призвело до кращої відповідності людському зору (зокрема, покращило чутливість до форми порівняно з текстурою) та підвищеної робастності моделі. ViT-22B стала новим стандартом для універсальних моделей аналізу зображень. Її успішно інтегрували в мультимодальну систему PaLM-E, у якій поєднуючи LLM вдалось значно поліпшити результати в робототехніці.

Сегментація зображень і універсальні моделі. У 2023 р. компанія Meta створила універсальну модель сегментації Segment Anything Model (SAM), навчену на наборі даних SA-1B (1,1 млрд. масок та 11 млн. зображень) [12]. Архітектура SAM містить потужний ViT-кодувальник (версія ViT-Huge) для отримання ознак та спеціалізований декодер для масок [12, 13]. Основна властивість SAM полягає у здатності сегментувати будь-який об'єкт без донавчання (*Zero-Shot*) за допомогою «підказок» (точок, рамок, тексту). Завдяки цьому SAM досягає конкурентної якості порівняно з моделями, які повністю навчені на конкретні задачі [13]. До коду та самої моделі надано вільний доступ, що значно стимулювало дослідження універсальних сегментаційних підходів. Отримані результати показали, що навіть дуже великі ViT-моделі можна ефективно застосовувати для універсальної сегментації на основі «підказок».

Самонавчання та універсальні ознаки. Наступним кроком у створенні універсальних моделей стало представлення моделі DINOv2 від Meta [14], яка довела, що великі ViT можна ефективно навчати без учителя. DINOv2 поєднує низку прийомів масштабування даних та архітектурної оптимізації. Зокрема, ViT навчений із 1 млрд параметрів на великому (з різноманітних джерел) наборі зображень. DINOv2 без донавчання перевершила попередні методи (наприклад, OpenCLIP) на багатьох критеріях класифікації, детектування і сегментації. Виконавши відокремлення, отримали компактніші версії (мільйонів параметрів), які майже не поступаються оригінальній моделі за якістю. Цей результат підтверджує, що SSL разом із масштабними ViT може продукувати високоякісні універсальні ознаки.

Ефективні та компактні ViT. Незважаючи на успіхи великомасштабних ViT, залишається актуальною проблема ефективності на пристроях із обмеженими ресурсами або в реальному часі. У

праці [15] запропоновано кілька архітектур для прискорення ViT без значної втрати точності. Зокрема, EfficientViT – сімейство швидких моделей від Microsoft, які оптимізують операції з пам'яттю в шарі самоуваги. EfficientViT використовує «шарувату» будову блоків (лише один шар самоуваги між двома легкими шарами зворотного зв'язку) та каскадну групову увагу (*Cascaded Group Attention*), що зменшує дублювання обчислень. У результаті модель EfficientViT-M2 працює приблизно в 5,8 рази швидше, ніж модель MobileViT-XXS, перевершуючи її у точності на +1,8%.

Інший напрям – оптимізація самоуваги для мобільних пристроїв. Наприклад, у моделі MobileViT замінили мультиголову самоувагу на лінійну (*Separable Self-Attention*) зі складністю O(*k*) замість O(k^2), і досягли точності 75,6% на ImageNet із лише ~3 млн. параметрів [16]. Ця модель у 3,2 рази швидша, ніж попередня MobileViT, що робить її однією з найкращих для мобільних застосувань у 2023 р.

Таблиця 2

Таблиця характеристик моделей ViT (2023–2025)

Модель (рік)	**Кількість параметрів**	**Основні особливості та досягнення**
Swin Transformer V2 (2022) [7]	3 млрд.	Ієрархічний ViT з віконною самоувагою; покращена стабільність навчання на зображення високої роздільності; Top-1 ≈ 84% на ImageNet
ViT-22B (2023) [11]	22 млрд.	Найбільший щільний ViT; SOTA на багатьох задачах; покращена робастність і узгодженість із особливостями людського зору
Segment Anything (2023) [12, 13]	~600 млн.	Універсальна модель сегментації (ViT-H backbone), навчена на 1,1 млрд. масок; Zero-Shot сегментація будь-яких об'єктів
DINOv2 (2023) [14]	1 млрд.	Самонавчальний ViT для універсальних ознак; перевершує OpenCLIP у кількох критеріях; є відокремлені компактні версії
EfficientViT (2023) [15]	<10 млн. (M-серії)	Швидкий ViT із каскадною груповою увагою; ~5–7 раз прискорення над MobileViT за однакової або вищої точності
MobileViT v2 (2023) [16]	3 млн.	Полегшена лінійна самоувага; 75,6% Top-1 на ImageNet; у 3,2 рази швидша за MobileViT на мобільному пристрої
LaViT (CVPR 2024) [17]	~88 млн. (Base)	Повна самоувага лише в частині шарів; менше надмірних обчислень; точність на рівні або вище за стандартні ViT
big.LITTLE ViT (2024–2025) [18]	—	Два паралельних трансформери (великий і малий) для токенів різної важливості; суттєво менші обчислювальні витрати за високій точності

Нові напрямки (2024–2025). ViT продовжує інтенсивно розвиватися у напрямку скорочення обчислень у шарі самоуваги без втрати точності. Наприклад, модель LaViT (*Less-Attention Vision*

Transformer) виконує повну самоувагу лише в деяких шарах кожної стадії, а в решті шарів «перетво-рює» матриці уваги, обчислені раніше. Такий підхід дає змогу зменшити дублювання обчислень і уникнути насичення уваги (*Attention Saturation*) [17]. За результатами експериментів з моделлю LaViT досягли порівняної або вищої точності в задачах класифікації, детектування та сегментації за меншої обчислювальної вартості.

Інший цікавий підхід запропоновано в моделі big.LITTLE ViT на основі двох паралельних трансформерів (великий і малий), в які динамічно спрямовуються токени різної важливості [18]. Це дало змогу суттєво скоротити сумарні витрати, зберігаючи високу якість. Такий підхід узагальнює тенденцію 2024–2025 рр. – зробити моделі ViT придатнішими для широкого застосування, балансуючи між точністю та ефективністю.

Крос-модальне навчання з використанням ViT.

Об'єднання зображень та тексту (Contrastive Learning, CLIP). Значному прогресу сприяло поєднання крос-модального навчання та ViT із мовними моделями для навчання на парах «зображення – текст». Ідея полягає в тому, щоб навчити модель, яка могла б зіставляти зображення з відповідним описом (та навпаки) у спільному просторі ознак. Прикладом такої реалізації стала модель CLIP (*Contrastive Language-Image Pretraining*), запропонована у праці [4]. Тут використали два окремих кодувальники: ViT для зображень та трансформер (або інша мовна модель, подібна до BERT) для текстових підписів. Обидва кодувальники перетворюють свій вхід у вектори однакової розмірності – блоки зображення і тексту, після чого модель навчається збільшувати їхню взаємну подібність для «правильних» пар і зменшувати для «випадкових» невідповідних пар.

Для цього оптимізують контрастну функцію втрат на основі підходу InfoNCE. Для вибірки з N пар (i-е зображення та i-ий текст) спрощено її можна записати як:

$$L_{CLIP} = -\frac{1}{N}\sum_{i=0}^{N}\left[\log\frac{\exp\left(\frac{sim(v_i,t_i)}{\tau}\right)}{\sum_{j=1}^{N}\exp\left(\frac{sim(v_i,t_j)}{\tau}\right)} + \log\frac{\exp\left(\frac{sim(v_i,t_i)}{\tau}\right)}{\sum_{j=1}^{N}\exp\left(\frac{sim(v_j,t_i)}{\tau}\right)}\right], \quad (4)$$

де $sim(v,t)$ – косинусна схожість між блоком зображення v та тексту t, а τ – температурний гіперпараметр. Перший доданок максимізує оцінку правильного тексту для заданого зображення проти інших текстів, другий – навпаки, правильного зображення для тексту. Таким чином, модель навчається «підтягувати» відповідні зображення і описи одне до одного у прихованому просторі, і «віддаляти» невідповідні.

Завдяки цьому підходу CLIP змогла навчатися на величезному масиві даних з інтернету – 400 млн. пар зображень з текстовими підписами, не потребуючи ручного маркування класів. Результати виявилися вражаючими: у режимі без донавчання на тестовому наборі CLIP досягла 76,2% точності на ImageNet. Таку ж точність отримували з класичним ResNet-50, навченим з учителем на ImageNet, але CLIP не використала жодного маркованого вручну класу ImageNet під час свого навчання. Отже, CLIP дає можливість класифікувати зображення без жодної навчальної вибірки за заданими класами, для цього достатньо лише текстового опису класів. Окрім того, її якість на інших наборах (наприклад, OCR-набір або далекий ImageNet-A) теж значно перевершила попередні підходи *zero-shot*. Це свідчить про те, що крос-модальний підхід може вчити моделі узагальненим уявленням, придатним для багатьох задач.

На рис. 2 подано спрощену архітектуру CLIP (зображення проходить через ViT-кодувальник, текст – через трансформер-кодувальник; отримані ознаки порівнюються попарно), а на рис. 3 показано функцію втрат InfoNCE під час навчання, яка прямує до мінімуму, якщо модель успішно навчається вирізняти «правильні» пари. Також автори порівняли CLIP з моделлю, яка намагається передбачати підписи (як мовну модель) – виявилося, що контрастний підхід навчається значно

ефективніше. Це узгоджується із загальною тенденцією високої ефективності контрастного навчання в самонавчанні для зображень.

Заслуговує на увагу також модель ALIGN [11], близька за ідеєю до CLIP, яка навчалася на ще більшому наборі даних (близько 1,8 млрд. пар). Обидві моделі підтвердили, що текстова анотація з мережі – потужне джерело сигналу для навчання зорових моделей.

Модель ViLBERT [14] була ранньою спробою об'єднати CNN-візуальні ознаки і BERT у спільній архітектурі, хоча вона ще спиралася на CNN для виділення ознак зображення.

Моделі CLIP та ALIGN започаткували розвиток нової генерації візуально-лінгвістичних трансформерів: від двохпотокових (*Dual-Stream*) до однопотокових (*Single-Stream*) моделей, де зображення і текст об'єднуються на рівні спільного трансформера. На сьогодні дослідники прагнуть повністю усунути потребу в окремих згорткових екстракторах – використовувати лише трансфор-мери для всіх модальностей.

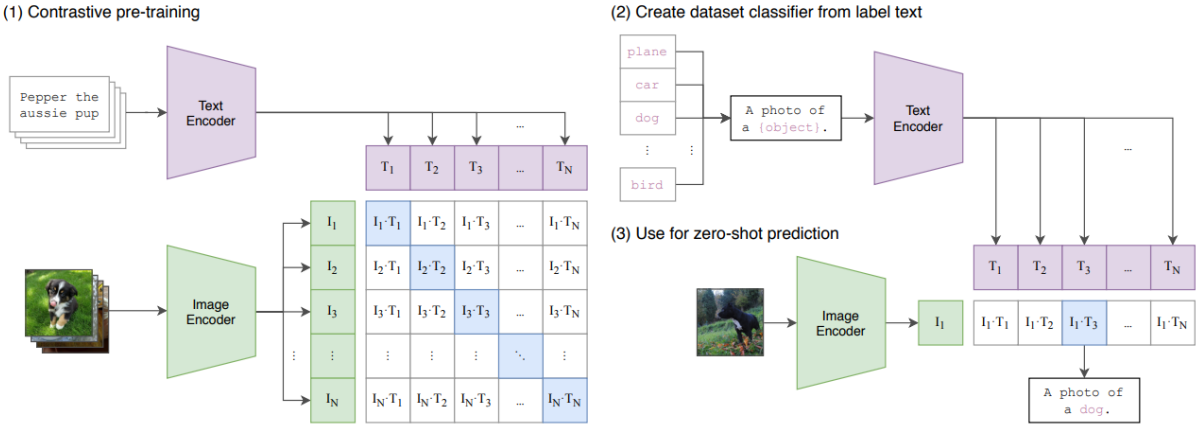

Рис. 2. Архітектура CLIP.

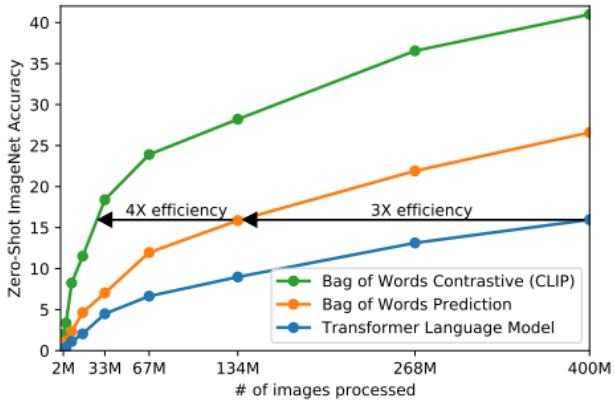

Рис. 3. Залежність точності на ImageNet у режимі *zero-shot* від кількості оброблених зображень для трьох підходів.

Розширення на інші модальності: відео, аудіо та сенсорні дані. Крос-модальне навчання не обмежується парою «зображення – текст». Сучасні системи можуть одночасно обробляти відеоряди, звук, текстові підписи, дані від сенсорів тощо. ViT активно застосовують як компонент таких мультимодальних моделей. Наприклад, Video-Audio-Text Transformer (VATT) [19], використовує єдину архітектуру трансформера одночасно для трьох модальностей: відео (послідовність кадрів), аудіо (звукоряд) і текст (рис. 5). VATT отримує на вході необроблені сигнали (пікселі відео, аудіохвилю, текстові набори) і навчається в режимі самонавчання, використовуючи мультимодальні контрастні втрати – подібно до згаданого вище принципу CLIP, але узагальнено на випадок трьох різнорідних входів.

Цікаво, що автори VATT навіть експериментували з універсальною моделлю – одним трансформером із поділеними вагами між усіма модальностями, тобто ті самі параметри обробляли і зображення, і звук, і текст. Хоча така уніфікація складна, вона демонструє прагнення до спільного простору представлень для всього, чого мозок людини досягає природно.

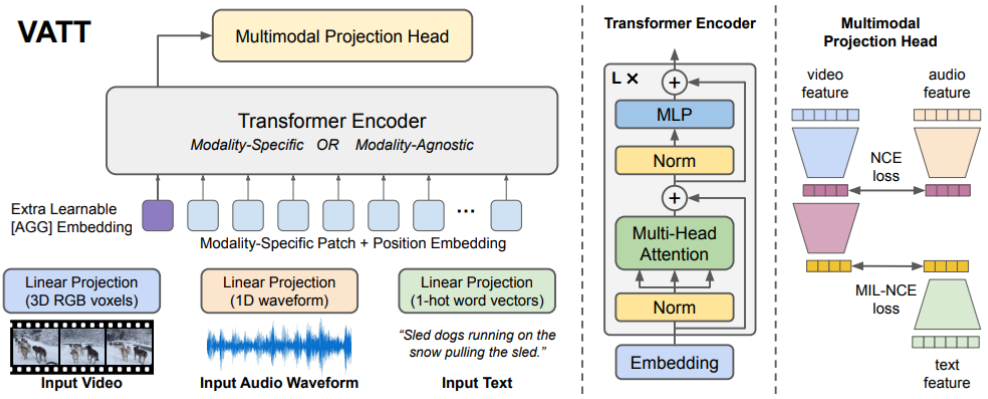

Рис. 4. Архітектура VATT (Vision, Audio, and Text Transformer)

Візуальний трансформер VATT із самонавчанням досягнув точності 82,1% на Kinetics-400 та 72,7% на Kinetics-700. Також відеотрансформер, донавчений для статичних зображень, показав точність 78,7% на ImageNet. Аудіо-трансформер VATT встановив рекорд розпізнавання звукових подій на основі хвилі, досягнувши mAP 39,4% на AudioSet. Трансформери довели універсальність та переваги самонавчання через крос-модальне навчання.

Отже, ViT виявились гнучкими для інтеграції з мовними моделями, аудіомодальностями та сенсорними потоками, використовуючи механізм уваги для крос-модального злиття інформації.

Самонавчання для ViT.

Важливість самонавчання для ViT. ViT мають значний потенціал для побудови моделей, але для його реалізації потрібні великі обсяги даних. Спочатку для вирішення цієї проблеми використовували зовнішні колекції, наприклад, JFT-300M [2]). Однак такі великі марковані набори даних не завжди доступні. Тоді у нагоді стає SSL, де модель навчається на немаркованих даних, вирішуючи допоміжні задачі *(Pretext Tasks*). Для ViT це дає змогу спочатку попередньо натренувати транс-формер на великій множині зображень чи відео без маркерів, а потім здійснити тонке налаштування (*Fine-Tune*) на меншому обсязі маркованих даних. Це значно покращує узагальнення та стійкість.

Існують три основні категорії SSL у комп'ютерному зорі:

- ***контрастне навчання;*** модель навчається так, щоб схожі за змістом приклади розташовувалися ближче в прихованому просторі, а несхожі – далі. Прикладом можуть бути моделі SimCLR, MoCo, BYOL та ін., де «правильні» пари утворюються з різних доповнень одного зображення, а «випадкові» – з інших зображень; формально використовується функція втрат на зразок InfoNCE (аналогічна до CLIP, але для різних доповнень одного зображення замість «зображення – текст»).

- ***генеративне навчання*;** модель намагається відновити вихідні дані з їх спотвореної версії; класичні приклади включають автокодувальники GAN, VAE; для ViT набули популярності масковані автокодувальники (*Masked Autoencoders, MAE*), коли випадкові сегменти зображення маскуються, і модель має їх відновити; наприклад, у праці [12] подано модель MAE: ViT-кодувальник обробляє близько 25% сегментів, а декодер відновлює решту 75%.

- ***прогнозоване навчання*;** модель прогнозує приховану частину входу, але не обов'язково повністю відтворює оригінальний сигнал; наприклад, у моделі Masked Image Modeling (MIM) за аналогією з BERT [5]: маскуються сегменти, і трансформер пророкує їх за високорівневими

ознаками чи кодами; такі методи більше зосереджуються на семантиці, ніж на точному відтворенні пікселів.

Для ViT самонавчання значно підвищує ефективність використання даних. Наприклад, у праці [13] показали, що SSL може сформувати у ViT «внутрішні» увагові карти, які збігаються з об'єктами на зображеннях без ручних маркерів. Це свідчить про виникнення семантично узагальнених представлень, подібно до того, як трансформери в обробці природньої мови вчаться складної семантики без прямої вказівки.

DINO: Самовідокремлення без маркерів. DINO (скор. «self DIstillation, NO labels») – це метод, який дає змогу тренувати потужні візуальні представлення без ручного маркування, поєднуючи ідеї SSL, знаннєвого відокремлення (*Knowledge Distillation*) та використання ViT [20]. Суть полягає у підході «вчитель – учень», де обидві мережі мають однакову архітектуру, але з різними вагами: «учень» оновлюється класичним зворотним поширенням помилки, а «вчитель» отримує нові ваги через експоненційне ковзне середнє від «учня» і не має власних градієнтів. Завдяки цьому «вчитель», перевершуючи учня за точністю, залишається стабільнішим і поступово накопичує корис-ні ознаки та зазвичай застосовується у подальших прикладних задачах. Покращити узагальнення і чутливість до різних масштабів допомагає мультикультурний (*Multi-Crop*) метод: «вчитель» отримує глобальні сегменти (наприклад, 224×224), а «учень» – кілька локальних (наприклад, 96×96). Така схема змушує модель виявляти ознаки, стійкі до змін розміру та контексту, що суттєво підвищує якість самонавчання. DINO уникає колапсу, коли модель видає майже однакові вектори для будь-якого входу.

Відмінність від класичного відокремлення полягає в тому, що обидві мережі навчаються одночасно, маючи різні огляди того самого зображення. Завданням є наближення виходів «учня» до виходів «вчителя» за допомогою перехресної ентропії як функції втрат, незалежно від застосованих доповнень. Це надає моделі стійкості до варіацій без потреби в ручних маркерах. У результаті DINO показує чудові результати самонавчання на ImageNet і багатьох інших наборах, де попередньо натренована на ImageNet модель дає високу точність у лінійних і k-NN класифікаторах, а карти самоуваги у ViT виявляють об'єкти на зображеннях без жодного маркера, наближаючись до «вільної» сегментації. Додаткові дослідження щодо зміни розміру сегментів, EMA-оновлення чи масштабу (*Batch Size*) підтвердили гнучкість і масштабованість цього підходу.

Математичні аспекти SSL у ViT. Як приклад контрастного SSL, розглянемо формулу InfoNCE докладніше, оскільки вона стала базою для багатьох методів.

Нехай маємо блок з N зображень. Для кожного i-го зображення визначимо «правильну» пару як два різні доповнення x_i та $x_i^{'}$ цього зображення, які, пройшовши через модель (наприклад, ViT), дають вбудовані блоки h_i та $h_i^{'}$. «Випадковими» прикладами для h_i є блоки інших зображень h_j для $i \neq j$. Тоді функцію втрат для i-ої пари можна записати у вигляді

$$L_i^{contrast} = -\log \frac{\exp\left(\frac{sim(h_i, h_i^{'})}{\tau}\right)}{\exp\left(\frac{sim(h_i, h_i^{'})}{\tau}\right) + \sum_{i \neq j} \exp\left(\frac{sim(h_i, h_j^{'})}{\tau}\right)}, \tag{5}$$

де $sim(u, v)$ – косинусна подібність, τ – температура.

Мінімізуючи $L_i^{contrast}$ за всіма i, ми «притягуємо» пари (i, i') («правильні») і «відштовхуємо» пари (i, j) («випадкові»). Для реалізації цього підходу у випадку ViT знадобилися певні модифікації, зокрема, велика пам'ять для зберігання багатьох «випадкових» пар або використання стоп-градієнтів для уникнення колапсу, коли модель робить всі вбудовані блоки однаковими.

Для генеративного SSL формули інші. Розглянемо масковане автокодувальне навчання (*Masked Autoencoder Learning*): маємо оригінальні сегменти $(z_1, z_2, ..., z_N)$ і маскуємо випадковий піднабір *M*, подаючи залишок $\{z_i | i \notin M\}$ у кодувальник. Декодер отримує приховані вектори кодувальника та позиційні ознаки для всіх сегментів і генерує реконструйовані сегменти $\hat{z_j}, j \in M$.

Функція втрат тут може бути просто середньоквадратичною похибкою:

$$L_{MAE} = \frac{1}{|M|} \sum_{j \in M} \left\| \hat{z_j} - z_j \right\|^2, \quad (6)$$

Мінімізуючи похибку, модель вчиться інтерполювати відсутні частини зображення. Після такого навчання кодувальник ViT можна використовувати самостійно – його виходи вже містять узагальнені ознаки зображення. Практика показала, що ViT, навчений методом MAE, під час налаш-тування на класифікацію часто перевершує навчений «з нуля» і навіть конкурує з контрастними методами, хоч і без жодного маркеру.

Комбінації підходів: від SLIP до CoCa. Різні підходи SSL можна поєднувати. Наприклад, SLIP (*Self-supervision meets Language-Image Pre-training*) об'єднує [21] контрастне навчання на парах «зображення – текст» (аналогічно до CLIP [4]) та контрастне навчання лише на зображеннях (як SimCLR [3]). Автори SLIP зазначають, що таке подвійне контрастне навчання дає лише незначний приріст порівняно з чистим CLIP. Вони припускають, що обидві контрастні функції втрат не додають нової інформації.

Натомість CoCa (*Contrastive Captioner*) поєднує дві функції втрат – контрастну і генеративну, навчаючи модель одночасно [22] відрізняти потрібний текст від неправильного для зображення (контрастно) і відновлювати коректний текст за зображенням (генеративно). Такий підхід дає можливість отримати ширші уявлення, оскільки генеративні методи краще зберігають інформацію. Сьогодні багато мультимодальних моделей включають завдання маскування/реконструкції (*Masking/Reconstruction*) для покращення якості представлень.

Результати та обговорення

Аналіз літературних джерел свідчить, що ViT разом із підходами крос-модального та самонавчання здатні вивчати узагальнені уявлення про дані з різних джерел – від зображень та тексту до аудіо й сенсорних сигналів. Однак ще зберігається низка обмежень, які стримують застосування ViT у реальних проєктах, особливо коли йдеться про масштабованість, інтерпретованість та ефективність обробки мультимодальних даних.

Основні виклики та обмеження існуючих методів.

*Обчислювальні труднощі та масштабованість***.** Одна з найважливіших перешкод у крос-модальному навчанні на базі ViT полягає у значних обчислювальних витратах. Механізм самоуваги має квадратичну складність відносно кількості токенів [2, 5], що ускладнює роботу з великими масивами мультимодальних даних (зображення + текст, зображення + аудіо, або навіть тримодальні поєднання).

Інтерпретованість та якість даних. У середовищі розгорнутих ViT критичною проблемою стає пояснюваність (*Explainability*). Коли модель обробляє одразу кілька типів сигналів (зображення, текст, аудіо), прості методи візуалізації (*Attention Maps*) або часткові метрики внеску окремих токенів не завжди дають змогу збагнути логіку роботи самої мережі [20]. У чутливих сферах, як медицина чи автономний транспорт, це породжує питання довіри до системи: користувачі прагнуть зрозуміти, чому модель ухвалила те чи інше рішення.

Не менш важливою перешкодою є *якість початкового набору даних*. У великих відкритих колекціях чимало дублікатів, спотворених прикладів та навіть токсичних матеріалів, що може спричиняти упередження в роботі моделей [11]. Водночас у високоспеціалізованих завданнях (наприклад, у медичних) доступ до якісної анотованої інформації ускладнений через політику конфіденційності чи відсутність публічних наборів даних [23]. Незважаючи на існування процедур

очищення й фільтрації даних, жодна з них не може повністю нейтралізувати всі помилки чи упередження.

Залежність від приростів та архітектурні обмеження. Більшість самонавчальних стратегій (контрастні, генеративні) покладаються на штучні прирости зображень або текстових описів, щоб модель навчилась інваріантності до несуттєвих змін [3, 21]. Хоча це позитивно впливає на узагальнюючу здатність, неправильний або надто агресивний набір приростів може призводити до викривлень у репрезентаціях, особливо в крос-модальних задачах, де потрібно синхронізовано трансформувати кілька видів сигналів.

Додатковий виклик пов'язаний з *архітектурою самих трансформерів*: квадратична складність механізму самоуваги стимулює пошук ефективніших або селективніших механізмів (*Sparse-Attention*) [17]. Деякі дослідницькі групи експериментують із гібридними структурами, зокрема, поєднують графові мережі та трансформери [24] чи біонатхненне масштабуванням. Проте жодне з цих рішень поки не стало універсальним стандартом, оскільки кожне оптимізоване переважно під певний тип задачі або набір даних і може втрачати ефективність в інших сценаріях.

Вирішення проблем та перспективи подальшого розвитку.

Поєднання кількох завдань у самонавчанні. Один із ключових шляхів зменшення залежності від великої кількості даних та складних приростів – одночасне навчання моделі на різних завданнях. Наприклад, ViT може паралельно виконувати масковану реконструкцію зображень, контрастне зіставлення та прогнозування елементів у даних [14, 22]. Якщо під час тренування він, з одного боку, намагається відновити зашумлені фрагменти картинки, а з другого – шукати текстовий опис, що найбільше відповідає цьому зображенню, модель отримує багатогранні та глибші уявлення про дані. Така багатозадачність допомагає краще узагальнювати знання та виявляти «зашумлені» чи неточні вхідні дані.

Легковагові та адаптивні моделі. Щоб зменшити обчислювальні витрати, активно досліджують кілька підходів: (*1*) оптимізація уваги; розріджена (*Sparse*) або ієрархічна самоувага дає змогу скоротити кількість обчислень та пам'яті, необхідних для роботи з великими вхідними обсягами [10, 18]; (*2*) адаптоване налаштування; пропонують вбудовувати невеликі адаптаційні блоки, які можна швидко змінювати й підлаштовувати до нових завдань або доменів [23]; (*3*) кілька етапні і самонавчальні методи; модель проходить основне навчання на великому, але необробленому наборі, а потім донавчається на дуже обмеженій вибірці з анотаціями [11]. Завдяки цим підходам ViT можна запускати навіть на мобільних пристроях або в системах, де дуже важливо швидко отримувати результат і немає можливості довго навчати гігантські мережі.

Інтеграція з великими мовними моделями. Поєднання ViT із LLM відкриває можливості для створення систем, що здатні працювати одразу з текстовою та візуальною інформацією, а також з аудіо чи іншими сенсорними сигналами [4]. Сучасні розробки, зокрема, ImageBind, пропонують передавати блоки зображень безпосередньо в LLM. У результаті мовна модель може не лише описувати об'єкти, а й давати логічні пояснення того, що відбувається на зображенні [25].

Висновки

Досліджено сучасні можливості трансформерів зору, зокрема їхню інтеграцію з крос-модальним підходом та самонавчальними методами. Проаналізовано основні архітектурні рішення й виклики, а також показано, що саме поєднання крос-модального навчання з самонавчанням дає змогу краще використовувати великі та різнорідні масиви даних.

Завдяки крос-модальності модель вчиться узгоджувати різні типи інформації, наприклад, текст і зображення або навіть аудіо та візуальний ряд. Це робить систему «розумнішою», адже вона може поєднувати контекст із декількох джерел, зменшуючи ймовірність поверхневих помилок та підвищуючи точність аналізу. Водночас самонавчання сприяє залученню не тільки маркованих прикладів, а й величезних обсягів даних без будь-яких маркерів. Таким чином, модель отримує змогу виявляти загальні закономірності в необроблених вибірках, не обмежуючись стандартними наборами, які бувають невеликими й можуть містити шум чи упередження. Поєднання

самонавчання та крос-модальних підходів розв'язує низку проблем: економить ресурси на маркуванні даних; розширює сферу застосування; формує глибші репрезентації.

Отже, можна очікувати, що спільне використання ViT, самонавчання та крос-модального підходу й надалі буде одним із визначальних напрямів розвитку штучного інтелекту.

Перелік використаних джерел

[1] R. Szeliski, "Computer Vision: Algorithms and Applications", 2nd ed., Springer Cham, 2022, XXII, 925 p. https://doi.org/10.1007/978-3-030-34372-9

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale", 2021, arXiv preprint, arXiv:2010.11929. https://arxiv.org/abs/2010.11929

[3] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations", Proc. 37th Int. Conf. on Machine Learning (ICML), 2020, P. 1597–1607.

[4] A. Radford, J. W. Kim, C. Hallacy, and A. Ramesh, "Learning Transferable Visual Models from Natural Language Supervision", Proc. of the 38th Int. Conf. on Machine Learning (ICML), 2021, arXiv preprint, arXiv:2103.00020. https://arxiv.org/abs/2103.00020

[5] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention Is All You Need", NIPS'17: Proc. Of the 31st Int. Conf. on Neural Information Processing Systems, 2017, P. 6000–6010.

[6] H. Touvron, M. Cord, M. Douze, et al., "Training data-efficient image transformers & distillation through attention", Proc. 38th Int. Conf. on Machine Learning (PMLR), 2021, Vol. 139, P. 10347–10357.

[7] Z. Liu, Y. Lin, Y. Cao, and H. Hu, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows", 2021, arXiv preprint, arXiv:2103.14030. https://arxiv.org/pdf/2103.14030, https://doi.org/10.1109/ICCV48922.2021.00986

[8]. K. Han, A. Xiao, E. Wu, et al., "Transformer in Transformer", Advances in Neural Information Processing Systems (NeurIPS), 2021, Vol. 34. https://arxiv.org/abs/2103.00112

[9] W. Wang, E. Xie, X. Li, et al., "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions", Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV), 2021, P. 568–578. https://doi.org/10.1109/ICCV48922.2021.00061

[10] H. Wu, B. Xiao, N. Codella, et al., "CvT: Introducing Convolutions to Vision Transformers", Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV), 2021, P. 22–31. https://doi.org/10.1109/ICCV48922.2021.00009

[11] C. Jia, Yi. Yang, Ye Xia, et al., "Scaling Up Visual and Vision-Language Representation Learning with Noisy Text Supervision", Proc. of the 38th Int. Conf. on Machine Learning (PMLR) 2021, Vol. 139, P. 4904–4916.

[12] K. He, X. Chen, S. Xie, Yet al., "Masked Autoencoders Are Scalable Vision Learners", arXiv preprint, arXiv:2111.06377, 2022. https://arxiv.org/abs/2111.06377

[13] A. Kirillov, E. Mintun, N. Ravi, et al., "Segment Anything", Proc. of the IEEE/CVF Int. Conf. on Computer Vision (ICCV), 2023, https://doi.org/10.1109/ICCV51070.2023.00371

[14] M. Caron, H. Touvron, I. Misra et al., "DINOv2: Learning Robust Visual Features without Labels", arXiv preprint, arXiv:2304.07193, 2023. https://arxiv.org/abs/2304.07193

[15] X. Liu, H. Peng, N. Zheng, et al., "EfficientViT: Memory Efficient Vision Transformer with Cascaded Group Attention", Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2023, P. 14421–14430. https://arxiv.org/abs/2305.07027 https://doi.org/10.1109/CVPR52729.2023.01386

[16] S. Mehta, and M. Rastegari, "MobileViT v2: An Efficient Neural Architecture for Mobile Vision", 2023, arXiv preprint, arXiv:2206.02680. https://arxiv.org/abs/2206.02680

[17] Y. Zhang, X. Li, S. Lin, and K. He, "You Only Need Less Attention at Each Stage in Vision Transformers", Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2024. https://doi.org/10.1109/CVPR52733.2024.00579

[18] H. Guo, Yu. Wang, Z. Ye, Ji. Dai, and Yu. Xiong, "big.LITTLE ViT: Dynamic Token Routing for Vision Transformers", arXiv preprint, arXiv:2410.10267, 2024. https://arxiv.org/abs/2410.10267

[19] A. Akbari, L. Yuan, R. Qian, et al., "VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text", Proc of the 35th Int. Conf. on Advances in Neural Information Processing Systems (NeurIPS), 2021, arXiv preprint, arXiv:2104.11178.

[20] M. Caron, H. Touvron, I. Misra, et al., "Emerging Properties in Self-Supervised Vision Transformers", 2021, arXiv preprint, arXiv:2104.14294.

[21] K. Mu, R. Salakhutdinov, and H. Fan, "SLIP: Self-supervision meets Language-Image Pre-training", 2022, arXiv preprint, arXiv:2112.12750.

[22] Ji. Yu, Z. Wang, V. Vasudevan, et al., "CoCa: Contrastive Captioners are Image-Text Foundation Models", 2022, arXiv preprint, arXiv:2205.01938.

[23] R. Azad, A. Kazerouni, M. Heidari, et al., "Advances in medical image analysis with vision Transformers: A comprehensive review", Medical Image Analysis, 2024, 91, 103000. https://doi.org/10.1016/j.media.2023.103000

[24] D. Shan, and G. Chen, "GvT: A Graph-based Vision Transformer with Talking-Heads Utilizing Sparsity, Trained from Scratch on Small Datasets", 2024, arXiv preprint, arXiv:2404.04924.

[25] R. Girdhar, A. El-Nouby, Z. Liu, et al., "ImageBind: One Embedding Space to Bind Them All", Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2023, arXiv preprint, arXiv:2305.05665. https://doi.org/10.1109/CVPR52729.2023.01457

Olena Stankevych[1], Danylo Matviikiv[2]

[1] Department of Computer Design Systems, Lviv Polytechnic National University,
S. Bandery st., 12, Lviv, Ukraine, E-mail: olena.m.stankevych@lpnu.ua, ORCID 0000-0002-5977-6351
[2] Department of Computer Design Systems, Lviv Polytechnic National University,
S. Bandery st., 12, Lviv, Ukraine, E-mail: danylo.o.matviikiv@lpnu.ua, ORCID 0009-0009-2643-2809

SELF-SUPERVISED VISION TRANSFORMERS FOR CROSS-MODAL LEARNING (REVIEW)

Received: February 28, 2025 / Revised: March 28, 2025 / Accepted: April 01, 2025

Abstract. Computer vision systems are increasingly expanding their application in visual data analysis. Model training methods are undergoing the greatest development and improvement as the results of this stage significantly affect the final classification of objects and the interpretation of input information.

Typically, computer vision systems use convolutional neural networks for training (Convolution Neural Network, CNN). The disadvantages of such systems are significant limitations in cross-modal learning, multimodality implementation, labeling of large amounts of data, etc. One of the ways to overcome these problems is to use Vision Transformers (ViT), which, compared to classical CNNs, have higher performance due to reduced inductive biases and high parallel computing efficiency. Introducing Self-Supervised Learning (SSL) technologies can significantly reduce the dependence on manually labeled data, contributing to the formation of generalized representations of images. Cross-Modal Learning (CML) expands the possibilities of processing them by combining data of different types. The development of the new approach, combined with the capabilities of cross-modal learning and self-learning in ViT in a single architecture, will ensure adaptability, efficiency, and system scalability in various applications.

The research aims to provide a detailed overview of ViTs, approaches to their architecture, and methods for improving their efficiency. The mathematical foundations of the key concepts of ViT, cross-modal learning and self-learning, the main modifications of ViT, and their integration with SSL and CML technologies are considered. A comparison of methods using characteristics, performance, and efficiency is provided. The key challenges and prospects facing researchers and developers while creating universal models in computer vision are outlined.

ViTs change computer vision by capturing global dependencies on images. Despite some challenges, ViTs provide excellent scalability and performance for large datasets. The active search for methods to overcome their limitations makes ViTs a key tool for improving image classification, object detection, and other computer vision tasks.

Keywords: Vision Transformers; Self-Supervised Learning; Cross-Modal Learning; Computer Vision; Deep Learning