DOI: https://doi.org/10.36910/6775-2524-0560-2024-54-17
УДК 004.85
Недашківський Богдан Миколайович, аспірант
https://orcid.org/0000-0002-9886-2674
Національний університет водного господарства та природокористування, м. Рівне, Україна

МЕТОДИ РОЗПІЗНАВАННЯ ТА ОБРОБКИ ЗОБРАЖЕНЬ ЗАДОПОМОГОЮ ЗОРОВОГО ТРАНСФОРМЕРА

Недашківський Б.М. Методи розпізнавання та обробки зображень задопомогою зорового трансформера. У цій роботі основна увага зосереджена на використанні можливостей Зорового Трансформера (ViT) як основи для досліджень розпізнавання та обробки зображень. Використання даної архітектури мотивується її вмінням моделювати довгострокові залежності, таким чином долаючи обмеження, пов'язані зі згортковими нейронними мережами (CNN), які обмежені локальними рецептивними полями. Не зважаючи на ефективність зорового трансформера у зборі глобальної інформації, його виняткова залежність від таких даних є неоптимальною для сценаріїв із зображеннями з кількома мітками. Ці зображення за своєю суттю містять різноманітні об'єкти, що охоплюють різні категорії, масштаби та просторові відносини. У світлі цього дослідження визнається неефективність покладатися виключно на глобальну інформацію для ефективної обробкитаких комплексних візуальних даних. Дослідження спрямоване на усунення цього обмеження шляхом дослідження стратегії, яка доповнює модель ViT додатковим механізмом, здатним включати контекстну інформацію, що стосується зображень із кількома мітками, завдяки інтеграції декількох многовидів та їх злиттю у єдиний многовид. Мета полягає в тому, щоб підвищити здатність моделі розрізняти та розпізнавати об'єкти, що характеризуються різноманітними атрибутами, розмірами та просторовим розташуванням. З'ясовуючи необхідність нюансованого підходу до вирішення проблем, пов'язаних із зображеннями з кількома мітками, це дослідження є спробою зробити внесок у поточний дискурс щодо вдосконалення методологій розпізнавання та обробки зображень. Дослідження стратегій доповнення ViT контекстно-залежними механізмами підкреслює прагнення до вдосконалення можливостей моделей на основі зору для більш надійних і універсальних застосувань у сфері комп'ютерного зору.
Ключові слова: зоровий трансформер, багатошарові зображення,моделювання, згорткові нейронні мережі, просторові відносини, механізми.

Nedashkivskyi B. Methods of Image Recognition and Processing Using TheVision Transformer. In this study, the primary focus is on leveraging the inherent capabilities of the pure Vision Transformer (ViT) as a foundational framework for research in image recognition and processing. The utilization of Transformer architecture is motivated by its proficiency in modeling long-range dependencies, thereby overcoming the limitations associated with Convolutional Neural Networks (CNNs), which are constrained by local receptive fields. Despite the efficacy of ViT in capturing global information, its exclusive reliance on such data proves suboptimal for scenarios involving multi-label images. These images inherently comprise diverse objects spanning various categories, scales, and spatial relations. In light of this, the study acknowledges the inadequacy of relying solely on global information for effective processing of such complex visual data. The research aims to address this limitation by investigating strategies that augment the ViT model with additional mechanisms capable of incorporatingcontextual information pertinent to multi-label images. The objective is to enhance the model's capacity to discern and recognize objects characterized by diverse attributes, dimensions, and spatial arrangements. By elucidating the need for a nuanced approach to address the challenges posed by multi-label images, this study endeavors to contribute to the ongoing discourse on advancing image recognition and processing methodologies. Theexploration of strategies to complement ViT with context-aware mechanisms underscores a commitment to refining the capabilities of vision-based models for more robust and versatile applications in the realm of computer vision.
Key words: vision transformer, multi-label images, long-range dependency modeling, convolutional neural networks, spatial relations, context-awaremechanisms.

Вступ та постановка проблеми. Розпізнавання зображень з кількома мітками є складною, але важливою задачею комп'ютерного зору, що представляє унікальні проблеми порівняно з розпізнаванням з однією міткою. Ключова перешкода полягає в тому, щоб точно ідентифікувати об'єкти з різними категоріями, масштабами та просторовими розташуваннями на зображенні. Це завдання має значну актуальність у практичних застосуваннях, таких як автономне водіння, мультимодальний аналіз і розпізнавання атрибутів людини. Ранні дослідження показали, що, незважаючи на надійні можливості глобального представлення згорткових нейронних мереж для зображень, моделі, попередньо навчені на наборах даних з однією міткою, можуть не бути оптимальними для даних завдань.

Трансформер – це архітектура глибокого навчання, заснована на механізмі, відомому як багатоголова самоувага (Multi-head attention). Він не має повторюваних одиниць, і тому вимагає менше часу на навчання, ніж попередні повторювані нейронні архітектури, такі як довготривала короткочасна пам'ять (LSTM). Вхідний текст розбивається на n-грами, закодовані як токени, і кожен токен перетворюється на вектор за допомогою пошуку з таблиці вбудовування слів. На кожному рівні кожен маркер контекстуалізується в межах контекстного вікна з іншими

(незамаскованими)маркерами за допомогою паралельного механізму багатоголової самоуваги, що дозволяє посилити сигнал для ключових маркерів і мінімізувати менш важливі маркери. Останніми роками архітектура зорового трансформера, спочатку була популяризована в задачах обробки природної мови, набула значного поширення в спільноті комп'ютерного зору. Успіх зорового трансформера складно переоцінити, оскільки він використовує механізм самоуваги (self-attention) для ефективного захоплення далеких залежностей між різними регіонами у вхідних зображеннях. Ця трансформаційна здатність знайшла широке застосування в таких завданнях, як класифікація зображень, знаменуючи зміну парадигми в підходах до розпізнавання зображень.

Незважаючи на те, що глибокі нейронні мережі досягли безпрецедентних досягнень у вилученні ознак і аналізі шаблонів розпізнавання, їхня недостатня прозорість створює проблеми для додатків, які вимагають відстежуваних і зрозумілих рішень. Щоб вирішити цю проблему, дослідники активно досліджують різні підходи до покращенняінтерпретації, прокладаючи шлях до прогресу в пояснюваному штучному інтелекті. Ця конвергенція проблем, пов'язаних з розпізнаванням та обробкою, а також пошуків інтерпретованих глибоких нейронних мереж закладає основу для всебічного дослідження розпізнавання та обробки зображень за допомогою зорового трансформера.

Аналіз останніх досліджень і публікацій. Наукові діячі сьогодення внесли значний вклад у розробку розпізнавання та обробки зображень з використанням зорового трансформера.

У дослідженні [1] було запропоновано зоровий трансформер у поєднанні з пірамідальною архітектурою, використовуючи метод поділу-перетворення-злиття, щоб запропонувати груповий кодувальник, якийвикористовує груповий кодер у трансформері, для отримання більшоїкількості можливостей. Крім того всі шляхи мали однакову топологію в кожному агрегованому трансформері. Також було наголошено, що експериментальні результати можна покращити за рахунок удосконалення патча, механізму уваги та інших частин трансформера. Було показано, що кількість гілок шифратора є конкретною вимірюваною величиною. Виконувалися завдання класифікації зображень на наборі даних CIFAR-10 і виявлення об'єктів на наборі даних COCO 2017.

Робота [2] присвячена дослідженню ефективності моделі для розпізнавання зображень з кількома мітками з урахуванням екземплярів. Суть методу полягає в отриманні карт уваги з урахуванням екземплярів, які використовуються для створення прогнозних пропозицій за допомогою методу динамічної локалізації екземплярів. Пропозиції передбачення дозволяють вибирати локальні області екземплярів на необробленому зображенні для інтегрованого навчання з кількома мітками. Щоб досягти метиусвідомлення екземплярів, весь процес було розділено на два етапи. По-перше, було лінійно зіставлено високовимірні патч-токени зорового трансформера з низьковимірними картами функцій із семантичною усвідомленістю категорій. По-друге, була запропонована стратегія обмежень для посилення обмежень просторового відношення до низьковимірних особливостей через механізм self-attention, щоб отримати карти уваги з урахуванням екземплярів. Що стосується динамічної локалізації екземпляра, було використано підхід, заснований на локалізації слабоконтрольованого об'єкта, для отримання локальних функцій регіону екземплярів, які разом із глобальними функціями утворюють конвеєр із кількома видами.

Крім того, варто зазначити праці наступних науковців: Крім того, варто зазначити праці наступних науковців: Чжан Чжімін, Лей Чженьюй, Омура Масаакі, Хасегава Хідеюкі, Гао Шанце [3], Ямабе Това, Сайто Такеші [4], Мен Лінчень, Лі Хендуо, Чень Бор-Чунь, Лань Шії, У Цзусюань, Цзян Лім Сер-Нам [5], Ван Пінпін, Чжан Сіньї, Чжао Юянь, Лі Юеті, Сюй Кайшен, Чжао Шуайінь [6], Фу Імінь, Лі Чжун, Чжан Цзовей [7], Фі Х., Чан Нам [8], Чен Гуантао, Чжоу Яньцун, Гао Шань, Лі Ін'юй, Ю Хао [9], Ху Юбін, Чен Юнь, Лу Аньці, Цао Чжицян, Вей Давей, Лю Цзе, Лі Чжицзюнь [10], Константинідіс Димитріос, Папастратіс Іліас, Димитропулос Космас, Дарас Петрос [11], Бо Ян, Ван Сівей, Чжу Ень, Лю Сіньван, Чень Вей [12], Цзюй Жуй-Ян, Лінь Тін-Ю, Чан Жень-Шіун, Цзянь Цзя-Хао, Лінь Юй-Шіан, Хуан Лю-Жуй-І [13], Албайрак Абдулкадір [14], Чень Хевей, Сян Чен, Цю Дун, Хуан Сюсян [15] та інших.

Проте, беручи до уваги вище зазначену наукову документацію, питання, пов'язане з розробкою технологій розпізнавання та обробки зображень з використанням зорового трансформера, все ще залишається недостатньо дослідженим та потребує подальшого опрацювання.

Постановка завдання. Метою роботи є дослідження технологій розпізнавання та обробки зображень з використанням зорового трансформера.

Викладення основного матеріалу дослідження. Функція зорового трансформера полягає в розбиванні зображеннь на послідовності патчів у виглядівхідних даних, а потім у використанні

багатьох складених блоків самоуваги (MSA) і прямої мережі (FFN) для моделювання довгострокових залежностей між ними. Формально для кожного вхідного зображення $I^{C\times H\times W}$, трансформерспочатку розбивається на 2D патчі з фіксованим розміром $X = [x_1, x_2, \dots, x_N]$, деN – кількість патчів, C, H і W позначають відповідно канал, висоту та ширину вхідного зображення. Потім ці патчі відображаються на D-вимірнівбудовування патчів $Z = [z_1, z_2, \dots, z_N]$ з лінійним шаром. Згодом маркер класу z_{cls}, додається до маркерів, слугуючи представленням всього зображення.

Позиційне вкладення E_{pos} також додається до цих токенів, щоб покращити їхпозиційну інформацію. Таким чином, послідовність токенів, введених у модель трансформера, представлена таким чином:

$$Z = [z_{cls};\ z_1,\ z_2,\ \dots,\ z_N] + E_{pos}, \tag{1}$$

де $z \in R^D$ та $E_{pos} \in R^{(N+1)\times D}$ відповідно.

Магістральна мережа моделі ViT складається з L-блоків, кожен з яких складається з MSA та FFN. Зокрема, l кодер в одній головці, послідовність токена Z_{11} проектується в матрицю запиту $Q_l \in R^{(N+1)\times D}$, ключову матрицю $K_l \in R^{(N+1)\times D}$ і матрицю значень $V_l \in R^{(N+1)\times D}$. Тоді матриця самоуваги $A_l \in R^{(N+1)\times(N+1)}$ обчислюється як:

$$A_l = ASoft\,max\left(\frac{Q_l K_l^T}{\sqrt{D}}\right)V_l = \left[a_{cls,l}; a_{1,l}, a_{2,l}, \dots, a_{N,l}\right]V_l \tag{2}$$

$A_{cls,l} \in R^{(N+1)}$ відома як матриця уваги класу, що відображає взаємодію між маркерами класу та іншими маркерами. Для більш ефективної уваги до різних підпросторів репрезентацій, матриця багатоголової уваги об'єднує вихідні дані кількох матриць одноголової самоуваги та проектує їх за допомогою матриці інших параметрів:

$$head_{i,l} = A(Z_l W_{i,l}^Q, Z_l W_{i,l}^K, Z_l W_{i,j}^V) \tag{3}$$

$$MSA(Z_l) = Concat(head_{i,l} \dots head_{H,l})W_l^O \tag{4}$$

де $W^Q_{i,l}$, $W^K_{i,l}$, $W^V_{i,l}$, W^O_l являють собою матриці параметрів у і-й голові уваги l-го блоку побудови, а Zlde-notes – вхідні дані в l блоці. Вихідні дані з MSA потім подаються в FFN для створення вихідних даних блоку збірки Z_{l+1}.Залишкові з'єднання також застосовуються до MSA та FFN наступним чином:

$$Z_l' = MSA(Z_l) + Z_l, Z_{l+1} = FFN(Z_l') + Z_l' \tag{5}$$

Остаточний прогноз створюється класифікатором, який бере маркер класу $z_{cls,L}$ з останнього блоку побудови як вхідні дані.

Хоча зоровий трансформер забезпечує найсучаснішу продуктивність у добре відомих наборах даних тестування завдяки його здатності моделювати коротко- та довгострокові зв'язки між різними областями зображення, він не має локальних індуктивних зміщень, притаманних згортковим нейронним мережам і неефективно моделює локальну інформацію. Задля подолання даної проблеми спрямованої на посилення моделювання локальної структури шляхом введення згорток у трансформер, перепроектування процесу токенізації латок і впровадження локальних механізмів уваги, або прийняття ієрархічних структур, подібних до CNN.

Однак більшість ViT працюють лише в евклідовому просторі значень інтенсивності пікселів, не звертаючи уваги на той факт, що представлення даних в інших різновидах може бути корисним для їх продуктивності. Крім того, карти уваги мають тенденцію бути подібними в глибших шарах ViT, і їхпредставлення перестає покращуватися.

Запропонований у даному дослідженні метод, заснований на механізмі мультимноговидової багатоголової уваги (MMA) може покращити продуктивність будь-якого ViT, замінивши його стандартний механізм самоуваги. Щоб досягти цього, MMA перетворює вхідну послідовність у точки у многовидах Евкліда, SPD і Грассмана та обчислює відстані між ними в цих многовидах, щоб створити об'єднану карту уваги з високою розрізнювальною здатністю. Далі детально описано кожне різноманіття, а також те, як обчислюються та об'єднуються окремі карти відстаней, щоб сформувати точну карту уваги.

1) Евклідовий многовид. За запитом $Q \in R^{L\times D}$ і ключовим $K \in R^{L\times d}$ векторами евклідова карта відстані MMA обчислюється подібно до стандартного ViT як:

$$D_E(Q, K) = \frac{QK^T}{\sqrt{d}} \tag{6}$$

Карта відстані $D_E \in R_{h\times L\times L}$, де h представляє кількість голів, виражає подібність між векторами запиту та ключами, причому більші значення позначають більшу відстань між двома векторами в евклідовому многовиді.

2) Многовид SPD. Являє собою особливий тип многовиду Рімана, який складається з точок, виражених у вигляді квадратних матриць M розміром d×d, і позначається так:

$$S_{++}^{d} = \left\{M \in R^{d\times d}: u^T M u > 0 \forall u \in R^d - \{0_d\}\right\} \quad (7)$$

Щоб матриця розглядалася як точка в многовиді SPD, вона повинна бути симетричною і мати сформовані власні значення. Коваріаційні матриці володіють такими властивостями, тому їх можна вважати точками в многовиді SPD. Таким чином, включення коваріаційних матриць у обчислення MMA є корисним для продуктивності ViT завдяки включенню додаткової інформації про вхідні дані, що підвищує рівень представлення вихідних ознак.

Враховуючи вектори запиту $Q \in R_{L\times d}$ і ключа $K \in R_{L\times d}$, початково використовується операція лінійної проекції, яка навчається, щоб зменшити розмірність векторів і підвищити обчислювальну ефективність запропонованого підходу. Зпрогнозований запит і ключові вектори можна визначити як Q_p, $K_p \in R_{L\times s}$, де s є розмірністю проекції. Після цього коваріаційні матриці цих векторів обчислюються у вигляді наступного співвідношення:

$$C_Q = cov(Q_p) = E\left[(Q_p - E[Q_p])(Q_p - E[Q_p])^T\right] \quad (8)$$

$$C_K = cov(K_p) = E\left[(K_p - E[K_p])(K_p - E[K_p])^T\right] \quad (9)$$

Завдяки своїм властивостям кожна з матриць $C_Q, C_K \in RL\times s\times s$ описує кластер L коваріаційних матриць, які лежать у точках на многовиді SPD. Потім відстань SPD між *i* коваріаційною матрицею запиту та *j* коваріаційною матрицею ключа обчислюється як:

$$D_{SPD}^{i,j}\left(C_Q^i, C_K^j\right) = \frac{\left\|C_Q^i, C_K^j\right\|_F}{s} \quad (10)$$

де $||\cdot||_F$ позначає норму Фробеніуса. Подібно до карти евклідової відстані, $D_{SPD} \in R^{h\times L\times L}$ виражає подібність між векторами запиту та ключовими векторами та кількісно визначає відстані між двома векторами в многовиді SPD.

3) Многовид Грассмана є також спеціальним типом многовиду Рімана, який вкладає всі p-вимірні лінійні підпростори, які лежать у d-вимірному евклідовому просторі. Многовид Грассмана, позначений як G(p,d), може бутипредставлений набором ортогональних матриць з ортогональної групи O(p) наступним чином:

$$G(p,d) = \frac{\{X\in R^{d\times p}: X^T X = I_p\}}{O(p)} \quad (11)$$

де X представляє будь-яку точку на многовиді Грассмана. Многовиди Грассмана зазвичай використовуються для моделювання послідовних і змінних у часі сигналів, оскільки будь-яка лінійна динамічна система може бути легко перетворена в точку у многовиді Грассмана. Як наслідок, перетворення вхідного простору в точки многовидid Грассмана може надати ViT додаткову інформацію щодо варіацій текстури та кольору наділянці зображення, що призводить до розширених представлень функцій із високою розрізнювальною силою.

4) Злиття многовидів. Після обчислення індивідуальних карт відстаней D_E, D_{SPD} і $D_G \in R_{h\times L\times L}$ у кожному многовидів пропонуються дві операції, позначені як раннє та пізнє злиття, для отримання представлення вихідних характеристик.

При операції раннього злиття три карти відстаней об'єднані разом, а потім використовується операція тривимірної згортки ϱ для вивченняпо-елементної вагової матриці для виконання ефективного відображення відстаней у різних різновидах і генерації вихідного представлення ознак.Точніше, три карти відстаней об'єднані разом, утворюючи карту:

$$D_{concat} = BN(concat(D_E, D_{SPD}, D_G)) \in R^{3\times L\times L} \quad (12)$$

з BN, що позначає пакетну нормалізацію. Після цього обчислюється вагова матриця, щоб об'єднати різні карти відстаней оптимальним чином:

$$W_D = soft\max\left(LeakyRELU(\varrho(D_{concat}))\right) \in R^{3\times L\times L} \quad (13)$$

Вагова W_D відповідає за обчислення відповідних ваг поелементно для точного об'єднання різних відстаней. Нарешті, представлення вихідної функції V′ обчислюється таким чином:

$$V' = soft\max(sum(W_D \otimes D_{concat}))\, V \in R^{L\times d} \quad (14)$$

де операція сумування (sum) застосовується до першого виміру добутку вагової матриці та

об'єднаних карт відстаней, фактично видаляючи перший вимір.

Пізнє злиття. У цій операції використовуються три паралельні механізми уваги, кожен з яких обробляє вхідні дані в різному многовиді, таким чином обчислюючи представлення характеристик многовиду:

V'_E = softmax(D_E)V, V'_{SPD} = softmax(D_{SPD})V і V'_G=softmax (D_G)V для евклідового, SPD та многовидів Грассмана відповідно. Тоді вихідна характеристика представлення $V' \in R^{L\times d}$ дорівнює:

$$V' = L\left(concat(V'_E, V'_{SPD}, V'_G)\right) \quad (15)$$

де L виконує лінійну проекцію від розміру представлення зчеплених ознак 3ddo кінцевого розміру представлення ознак d.

Рис. 1 – Зразки зображень із 8 класів набору даних ETH-80. Коженстовпець містить зображення з одного класу

Експериментальна оцінка запропонованого методу була виконана з використанням набору даних ETH-80, який містить зображення 8 різних класів (яблуко, автомобіль, корова, чашка, собака, кінь, груша та помідор), з 10об'єктами на клас (наприклад, 10 різних яблук) і 41 зображення на об'єкт, зроблене з різних 3D-ракурсів. Зразки зображень із набору даних показані на рисунку 1. В експерименті випадковим чином вибирали 5 зображень на об'єкт для навчання, а решту 36 використовували для тестування. Цей випадковий процес поділу повторювався 5 разів і повідомлялося про середню точність. Метою було передбачити правильний клас даного тестового зображення. Усі зображення ETH мають загальний розмір 256 × 256.

Для розрахунку дескрипторів HOG використовували 16×16-вимірні комірки, 9 бінів гістограми на комірку та 4-кратну блочну нормалізацію.Отриманий дескриптор HOG є 15 × 15 × 9 × 4 = 8100 – вимірним вектором із фіксованою нормою 225. Було також використано відсікання гістограми, де нормалізовані значення гістограми обрізалися на 0,2, а потім знову нормалізувалися.

Результати розпізнавання представлені в таблиці 1 для наступних методів:

1) Дескриптори HOG у SVM зі звичайним ядром RBF Евкліда Гауса.
2) Дескриптори HOG у SVM із геодезичним експоненціальним ядром.
3) Дескриптори коваріації в SVM з логарифмічним евклідовим ядром.
4) Класичний зоровий трансформер.
5) Запропонований зоровий трансформер з технологією многовидовоїсамоуваги.

Таблиця 1 – Точність класифікації в наборі даних ETH-80. Злиттямноговидів призводить до значного підвищення точності

Метод	Точність класифікації
HOG у SVM зі звичайним ядром RBF ЕвклідаГауса	89.49 ± 0.91
HOG у SVM із геодезичним експоненціальнимядром	90.45 ± 0.79
HOG у SVM із геодезичним експоненціальнимядром	88.94 ± 0.34
Класичний зоровий трансформер	91.09 ± 0.42
Зоровий трансформер з технологієюмноговидової самоуваги	93.06 ± 0.69

Для обчислення дескриптора регіону коваріації кожного зображення використовувався вектор ознак [x y I $|I_x|$ $|I_y|$ $|I_{xx}|$ $|I_{yy}|$], де x, y – розташування пікселів, I вказує значення інтенсивності, а I_x, I_y, ... – похідні. Матриціковаріації, обчислені на основі цих характеристик, являють собою матриці 7 × 7 SPD, які лежать на многовиді Sym+7.

Висновки. Стаття представляє нову техніку розпізнавання зображень, яка використовує архітектуру Visual Transformer (ViT). Запропонований метод відхиляється від стандартного підходу

ViT, запроваджуючи складний механізм для обробки складних взаємозв'язків у межах зображень. На відмінувід традиційного ViT, який покладається на багатосторонню самоувагу для моделювання довгострокових залежностей, техніка многовидової уваги привносить новий вимір у розуміння просторових зв'язків у зображеннях, завдяки чому модель демонструє підвищену продуктивність у захопленні складних функцій і тонкої контекстної інформації. Цей метод виявився особливо ефективним у зниженні обчислювальних витрат і підвищенні ефективності шляхом стратегічного зосередження на головних блоках.

Порівняно зі стандартним ViT, техніка многовидової уваги демонструє свою перевагу в обробці внутрішньої просторової надмірності в зображеннях.Можливість точного визначення та виділення відповідних регіонів значно сприяє точності та ефективності моделі.

Перспективами подальших досліджень є розробка більш ефективної методології впровадження зорового транформера у розпізнавання та обробку зображень різних класів.

Список бібліографічного опису

1. Ju R. Lin T., Chiang J., Jian J., Lin Y., Huang L. Aggregated Pyramid Vision Transformer: Split-transform-merge Strategy for Image Recognitionwithout Convolutions. 2022.
2. Hu Y., Jin X., Zhang Y., Hong H., Zhang J., Yan F., He Y., Xue H.. Diverse Instance Discovery: Vision-Transformer for Instance-Aware Multi-Label Image Recognition. 2022.
3. Zhang Z., Lei Z., Omura M., Hasegawa H., Gao S.. Dendritic Learning-Incorporated Vision Transformer for Image Recognition. *IEEE/CAA Journal of Automatica Sinica.* 2024. №11. P. 539-541. DOI:10.1109/JAS.2023.123978.
4. Yamabe T., Saitoh T. Vision Transformer-Based Bark Image Recognition for Tree Identification. 2023. DOI:10.1007/978-3-031-25825-1_37.
5. Meng L., Li H., Chen B., Lan S., Wu Z., Jiang Y., Lim S. AdaViT: Adaptive Vision Transformers for Efficient Image Recognition. 2021.
6. Wang P., Zhang X., Zhao Y., Li Y., Xu K., Zhao S. Analysis of blood cell image recognition methods based on improved CNN and Vision Transformer. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*. 2023. DOI:10.1587/transfun.2023EAP1056.
7. Fu Y., Li Z., Zhang Z. Progressive Learning Vision Transformer for Open Set Recognition of Fine-Grained Objects in Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*. 2023. P. 1-1. DOI:10.1109/TGRS.2023.3309091.
8. Phi H., Tran N. Evaluation Of Vision Transformer On Weather Image Recognition. *Tra Vihn University Journal Of Science*. 2023. DOI:10.35382/tvujs.13.3.2023.2431.
9. Cheng G., Zhou Y., Gao S., Li Y., Yu H. Convolution-Enhanced Vision Transformer Network for Smoke Recognition. *Fire Technology.* 2023. № 59. P. 1-24. DOI:10.1007/s10694-023-01378-8.
10. Hu Y., Cheng Y., Lu A., Cao Z., Wei D., Liu J., Li Z.. LF-ViT: Reducing Spatial Redundancy in Vision Transformer for Efficient ImageRecognition. 2024.
11. Konstantinidis D., Papastratis I., Dimitropoulos K., Daras P. Multi-Manifold Attention for Vision Transformers. *IEEE Access.* 2023. P. 12-15. DOI:10.1109/ACCESS.2023.3329952.
12. Bo Y., Wang S., Zhu E., Liu X., Chen W. Group-Attention Transformer for Fine-Grained Image Recognition. 2022. DOI:10.1007/978-3-031-06761-7_4.
13. Ju R., Lin T., Chiang J., Jian J., Lin Y., Huang L. Aggregated Pyramid Vision Transformer: Split-transform-merge Strategy for ImageRecognition without Convolutions. 2022.
14. Albayrak A.. Vision Transformer Based Photo Capturing System. *Balkan Journal of Electrical and Computer Engineering*. 2023. №11. DOI:10.17694/bajece.1345993.
15. Chen H., Xiang C., Qiu D., Huang X. Multicategory ImageRecognition Based on Image Semantic Features and Transformer. *Mobile Information Systems*. 2022. P. 1-8. DOI:10.1155/2022/4508507.

References

1. Ju R. Lin T., Chiang J., Jian J., Lin Y., Huang L. Aggregated Pyramid Vision Transformer: Split-transform-merge Strategy for Image Recognitionwithout Convolutions. 2022.
2. Hu Y., Jin X., Zhang Y., Hong H., Zhang J., Yan F., He Y., Xue H.. Diverse Instance Discovery: Vision-Transformer for Instance-Aware Multi-Label Image Recognition. 2022.
3. Zhang Z., Lei Z., Omura M., Hasegawa H., Gao S.. Dendritic Learning-Incorporated Vision Transformer for Image Recognition. *IEEE/CAA Journal of Automatica Sinica.* 2024. №11. P. 539-541. DOI:10.1109/JAS.2023.123978.
4. Yamabe T., Saitoh T. Vision Transformer-Based Bark Image Recognition for Tree Identification. 2023. DOI:10.1007/978-3-031-25825-1_37.
5. Meng L., Li H., Chen B., Lan S., Wu Z., Jiang Y., Lim S. AdaViT: Adaptive Vision Transformers for Efficient Image Recognition. 2021.
6. Wang P., Zhang X., Zhao Y., Li Y., Xu K., Zhao S. Analysis of blood cell image recognition methods based on improved CNN and Vision Transformer. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*. 2023. DOI:10.1587/transfun.2023EAP1056.
7. Fu Y., Li Z., Zhang Z. Progressive Learning Vision Transformer for Open Set Recognition of Fine-Grained Objects in Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*. 2023. P. 1-1. DOI:10.1109/TGRS.2023.3309091.
8. Phi H., Tran N. Evaluation Of Vision Transformer On Weather Image Recognition. *Tra Vihn University Journal Of*

Science. 2023. DOI:10.35382/tvujs.13.3.2023.2431.
9. Cheng G., Zhou Y., Gao S., Li Y., Yu H. Convolution-Enhanced Vision Transformer Network for Smoke Recognition. *Fire Technology.* 2023. № 59. P. 1-24. DOI:10.1007/s10694-023-01378-8.
10. Hu Y., Cheng Y., Lu A., Cao Z., Wei D., Liu J., Li Z.. LF-ViT: Reducing Spatial Redundancy in Vision Transformer for Efficient ImageRecognition. 2024.
11. Konstantinidis D., Papastratis I., Dimitropoulos K., Daras P. Multi-Manifold Attention for Vision Transformers. *IEEE Access.* 2023. P. 12-15. DOI:10.1109/ACCESS.2023.3329952.
12. Bo Y., Wang S., Zhu E., Liu X., Chen W. Group-Attention Transformer for Fine-Grained Image Recognition. 2022. DOI:10.1007/978-3-031-06761-7_4.
13. Ju R., Lin T., Chiang J., Jian J., Lin Y., Huang L. Aggregated Pyramid Vision Transformer: Split-transform-merge Strategy for ImageRecognition without Convolutions. 2022.
14. Albayrak A.. Vision Transformer Based Photo Capturing System.*Balkan Journal of Electrical and Computer Engineering*. 2023. №11. DOI:10.17694/bajece.1345993.
15. Chen H., Xiang C., Qiu D., Huang X. Multicategory ImageRecognition Based on Image Semantic Features and Transformer. *Mobile Information Systems*. 2022. P. 1-8. DOI:10.1155/2022/4508507.