ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2025. No. 1 (31)

УДК 004.85:004.94 DOI: https://doi.org/10.30837/2522-9818.2025.1.020

Д. Дашенков, К. Смеляков

РОЗШИРЕННЯ НАБОРУ ДАНИХ *IMAGENET* ДЛЯ МУЛЬТИМОДАЛЬНОГО НАВЧАННЯ З ТЕКСТОМ ТА ЗОБРАЖЕННЯМИ

Предмет дослідження: методи оброблення зображень для класифікації та інших завдань комп'ютерного зору з використанням мультимодальної інформації, зокрема текстових описів класів і зображень. **Мета статті –** розроблення мультимодального набору даних для класифікації зображень за допомогою аналізу текстової метаінформації. Отриманий набір має містити: дані зображень, класи зображень, а саме 1000 класів об'єктів, поданих на фото з набору *ImageNet*, текстові описи окремих зображень і текстові описи класів зображень загалом. **Завдання:** 1) на основі зображень набору *ImageNet* скомпілювати набір даних для навчання моделей-класифікаторів із текстовими описами класів зображень та окремих зображень; 2) на основі отриманого набору даних провести експеримент з навчання мовної нейронної мережі для підтвердження ефективності використання запропонованого підходу для виконання завдання класифікації. **Методи:** компіляція наборів даних вручну, навчання мовних нейронних мереж на основі архітектури *RoBERTa*. Навчання нейронної мережі проводилось за методом донавчання (f*ine-tuning*), а саме надбудови шару нейронної мережі на наявну модель для отримання нової моделі машинного навчання, здатної виконувати обране завдання. **Результати дослідження.** Створено набір даних, що комбінує дані зображень з текстовою інформацією. Отриманий набір даних є корисним для встановлення зв'язку між інформацією, яку модель машинного навчання здатна виокремити з фото, та інформацією, яку модель може виокремити з текстових даних. Мультимодальний підхід може застосовуватись у розв'язанні широкого спектра завдань, що продемонстровано на прикладі навчання мовної нейронної мережі. Навчена мовна модель обробляє опис зображень, що містяться в наборі даних, та прогнозує клас зображення, з яким пов'язаний цей опис. Модель покликана відфільтрувати нерелевантну текстову метаінформацію, покращуючи якість набору. **Висновки:** набори даних, які комбінують в собі декілька видів інформації, здатні надавати ширший контекст для розв'язання завдань, що, як правило, асоціюються лише з одним типом даних. Це дає змогу більш ефективно застосовувати методи машинного навчання.

Ключові слова: мультимодальне машинне навчання; класифікація зображень; оброблення природної мови; набори даних; текстова метаінформація.

Вступ

З-поміж завдань комп'ютерного зору класифікація зображень є однією з найбільш поширених та вирішуваних. Саме завдання визначається як вибір із двох (бінарна класифікація) або більше (мультикласова класифікація) класів, до яких належить певне зображення. У цій роботі розглядається завдання мультикласової класифікації незалежно від того, чи стоїть завдання вибору одного чи багатьох класів.

Класифікація зображень може виконувати декілька різних завдань предметної галузі. Найбільш поширене з них, що розглядається в цій роботі, – це розпізнавання певної кінцевої множини типів об'єктів на зображенні [1]. У цьому завданні типи об'єктів, які необхідно розпізнати, і є цільовими класами.

Особливість завдання класифікації полягає в існуванні більш релевантних і менш релевантних класів. Більш релевантні класи – це ті, для яких знаходження даних для тренування моделей нейронних мереж є простішим. Менш релевантні класи через брак даних для тренування становлять виклик для створення моделей з достатньо високою ефективністю. Для розв'язання цієї проблеми існує низка підходів:

– трансферне навчання дає змогу отримувати вищі показники ефективності роботи моделі на менш релевантних класах за допомогою навчання на більш релевантних класах [2];

– аргументація даних дає змогу створювати штучну, але близьку до реальної інформацію, що розширює вибірку для менш релевантних класів (наприклад, операції над зображеннями, зокрема кольорові зсуви, розмиття, перегортання, випадкова зміна розміру тощо) [3];

– мультимодальні підходи до навчання дають змогу отримувати інформацію про клас із сторонніх джерел унаслідок залучення даних з іншим типом (модальністю) [4].

У цьому дослідженні розглядається використання мультимодальних підходів для розв'язання завдання класифікації зображень. У цьому разі як другий тип

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 1 (31) ISSN 2524-2296 (online)

даних про зображення застосовується текст, що описує і самі зображення (так звані текстові анотації), і класи об'єктів, що необхідно розпізнати на поданих зображеннях. Інші підходи, зазначені вище, можуть використовуватись як разом із мультимодальним навчанням, так і окремо, проте це не є завданням цієї роботи.

Сучасні методи комп'ютерного зору значно розширюють свої можливості завдяки використанню мультимодальних підходів, що дають змогу інтегрувати різні типи даних, зокрема зображення та текстову інформацію [5]. Це відкриває нові перспективи для виконання таких завдань, як класифікація зображень завдяки поєднанню візуального контексту з текстовими описами, що підвищує точність і надійність моделей машинного навчання. Водночас розвиток цього напряму стикається з низкою викликів, серед яких: обмеження наявних наборів даних, недостатність текстових метаописів, що відповідають зображенням, та складність у навчанні моделей, здатних обробляти такі комплексні дані.

Для розв'язання окреслених завдань мультимодальне навчання пропонує інтеграцію методів оброблення природної мови (NLP) та комп'ютерного зору з метою створення моделей, здатних працювати із зображеннями й текстом як єдиним джерелом інформації. Текстові описи можуть доповнювати зображення, надаючи їм семантичну глибину, що полегшує інтерпретацію та класифікацію об'єктів. Особливу актуальність має розроблення якісних мультимодальних наборів даних, що об'єднують ці два типи інформації. Вони є ключовими для навчання моделей, які можуть ефективно використовувати зв'язки між текстом і візуальними даними [5, 6].

У статті зосереджено увагу на створенні мультимодального набору даних для класифікації зображень із застосуванням текстових описів класів та окремих зображень. Запропонований підхід спрямований на оптимізацію використання даних *ImageNet* [7] унаслідок їх інтеграції з текстовою метаінформацією, що забезпечує ширші можливості для аналізу та підвищує якість класифікації. Розроблений набір даних може стати основою для навчання мовних нейронних мереж, зокрема архітектур, подібних до *RoBERTa* [8], за допомогою донавчання, що забезпечить високу точність і гнучкість моделей.

Мультимодальні підходи демонструють значний потенціал у розв'язанні завдань, де традиційно використовуються лише окремі типи даних. Вони не лише розширюють сферу застосування методів машинного навчання, але й беруть до уваги додаткові контексти, необхідні для точного аналізу складної інформації.

Аналіз останніх досліджень і публікацій

Моделі машинного навчання є універсальним інструментом для розв'язання низки проблем комп'ютерного зору, зокрема сегментації зображень, оброблення та видозмінення зображень з метою досягнення бажаного естетичного ефекту, бінарної та багатокласової класифікації зображень, трекінгу об'єктів на відео, генерації зображень на відео або їх фрагментів тощо. Кілька останніх років методи, що впроваджують машинне навчання, набули значного розвитку. Архітектури моделей, що дають змогу працювати із зображеннями та відео, розвинулись від відносно простих згорткових нейронних мереж до резидуальних моделей, до рекурентних архітектур [9] та останнім часом до архітектур, що використовують механізми уваги, тобто трансформери та похідні від них моделі [10]. Оскільки методи трансформерів були запозичені зі сфери оброблення природної мови (NLP), застосування цих двох, на перший погляд, різних типів даних для досягнення кращих результатів обіцяє хороші результати за умови, що інформація з обох джерел може бути поєднана [10, 11].

Мультимодальний підхід до навчання й застосування нейронних мереж нині набув значного поширення. Зокрема передовою розробкою в цьому напрямі можна вважати модель CLIP. Зазначена модель є класифікатором об'єктів на зображеннях. Різниця CLIP від аналогів полягає в тому, що на основі текстової метаінформації модель здатна розширювати множину класів, які вона може ідентифікувати. Принцип роботи CLIP полягає у створенні спільного гіперпростору ембедингів із зображень та їх текстових анотацій. Під час навчання модель як вхід отримує і той, і інший тип даних. Завдання навчання – створити бінарний класифікатор, що відповідає на запитання "чи підходять одне одному зображення на текстовий опис?". Під час виконання модель, використовуючи заздалегідь підготовлені множини описів зображень

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2025. No. 1 (31)

і класів, намагається поставити відповідність між відомими класами та зображеннями. Для цього модель виконується багато разів – у парі з усіма підготовленими анотаціями до можливих класів зображень. Самі анотації є гіперпараметрами моделі. Результатом виконання алгоритму є множина результатів бінарних класифікацій. Далі, залежно від завдання, з цієї множини обирається один або декілька класів із найвищими показниками впевненості моделі. Модель успішно переноситься на більшість завдань і часто є конкурентоспроможною з повністю контрольованою базовою лінією без необхідності будь-якого спеціального навчання набору даних [12].

Мультимодальний підхід до навчання нейронних мереж використовується в низці особливих завдань, наприклад у виявленні чуток за допомогою інформації з текстових і візуальних даних. Найкритичніша складність у виявленні мультимодальних чуток полягає в захопленні як внутрішньомодальних, так і міжмодальних зв'язків із мультимодальних даних. Тоді як конвенційні методи зазвичай зосереджені на мультимодальному процесі синтезу, приділяючи мало уваги внутрішньомодальним відношенням. Новий метод, запропонований Л. Пенг, упроваджує глибоке метричне навчання, щоб ефективно отримувати мультимодальні зв'язки новин для виявлення чуток. Зокрема розроблено триплетний метод навчання на основі метрики, щоб виокремити внутрішньомодальні зв'язки між чутками та нечутками в кожному типі даних, а також порівняльне попарне навчання, щоб охопити міжмодальні зв'язки в мультимодальному режимі. Масштабні експерименти на двох реальних мультимодальних наборах даних доводять чудову ефективність цього методу виявлення чуток [13].

Навчання мультимодальних моделей зазнає негативних наслідків від таких проблем, як семантичні конфлікти, дублювання та шум. Хоча механізм уваги може бути використаний для часткового розв'язання окреслених проблем, він працює неявно й не може бути активно контрольованим. Більш перспективним методом розв'язання цього питання є інтеграція здатності міркувати до мультимодальних репрезентативних навчальних мереж [14].

Навчання зі слабким контролем продемонструвало потенціал у застосуванні корисних знань, прихованих за мультимодальними даними. Наприклад, маючи зображення та його опис, дуже ймовірно, що сегмент можна описати деякими словами в реченні. Хоча однозначна відповідність між ними повністю не відома, робота, запропонована А. Карпати на іншими [15], показує, що ці приховані зв'язки можна виявити за допомогою слабкоконтрольованого навчання. Потенційно більш перспективним застосуванням такого типу слабких методів нагляду є аналіз відео, де різні модальності, зокрема дії, аудіо, мови, приблизно вирівнюються на часовій шкалі.

Питання збору наборів даних актуальне не лише для зображень і тексту. Так, робота С. Шін та інших описує створення набору даних на основі відеокліпів корейської телереклами. Мітки та описи в цьому наборі даних містять різноманітну інформацію про контекст, наміри та емоції, що описується за допомогою зору та мови в кожному відеокліпі. Створений набір даних може розв'язати проблему відсутності публічної інформації для дослідження мультимодальної взаємодії з корейським. Очікується, що цей набір даних можна застосувати в конструкціях різних служб штучного інтелекту, таких як корейське оброблення діалогів, вилучення візуальної інформації та різні мультимодальні завдання аналізу інформації [16].

Важливим інструментом у підготовленні інформації для мультимодального навчання є визначення необхідних і достатніх характеристик в усіх типах даних. Метод, описаний в роботі Б. Чен та інших, демонструє можливість визначення таких характеристик за допомогою навчання моделей. Зазначений підхід демонструє ефективність використання так званого розплутування для вивчення подання високої вірогідності необхідності та достатності характеристик на основі мультимодального навчання. Зосереджуючись на необхідних і достатніх характеристиках та ігноруючи інші, модель вивчає більш інформативні та дискримінаційні подання, що призводить до кращої якості роботи моделі [17].

Отже, роль мультимодального навчання в розвитку методів машинного навчання полягає в інтеграції різних типів даних, таких як зображення, текст, відео, аудіо тощо, для отримання більшої кількості інформації про певний об'єкт чи явище без пошуку додаткових джерел однотипних даних.

Передові дослідження демонструють потенціал мультимодального підходу до навчання, даючи змогу створювати моделі, що, споживаючи більш різноманітну

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 1 (31) ISSN 2524-2296 (online)

інформацію з предметної галузі, показують кращі результати в деяких завданнях, ніж конвенційні моделі, що працюють з одним типом даних.

Дослідження, зосереджені на створенні мультимодальних наборів даних, розширюють для інших фахівців доступ до інформації з метою навчання. Водночас недоліком мультимодальних нейронних мереж є більший розмір мережі та більший обсяг даних. Методи визначення релевантної інформації та фільтрація нерелевантної допомагають зменшити вплив цього недоліку, зберігаючи якість роботи мережі.

Визначення не розв'язаних раніше частин загальної проблеми. Мета роботи, завдання

Нині завдання класифікації зображень (за критерієм наявності певного класу об'єктів на фото) є поширеним. Завдання класифікації зображень за фіксованого заданого набору класів із достатньою кількістю даних загалом є виконаним, адже існує безліч наявних якісних моделей машинного навчання, здатних опрацьовувати зображення з передбачуваними класами. Кінцевою метою нашого дослідження є створення методів класифікації зображень для випадків, коли не всі класи об'єктів відомі на момент навчання моделі машинного навчання або не всі відомі класи мають достатню кількість інформації для навчання моделі. Такі методи можуть спиратись на додаткові дані про об'єкти, які не доступні моделям машинного навчання, що обробляють тільки зображення, а саме текстову інформацію.

Для розроблення таких методів необхідно створити набір даних, що містять інформацію про зображення, про класи об'єктів на зображеннях і текстову метаінформацію для цих зображень. У цьому разі для отримання більш якісної та повної інформації про об'єкти на зображеннях необхідно додати в набір даних як текстові описи окремих зображень, так і описи самих класів об'єктів, їх зовнішнього вигляду, фізичних характеристик, зв'язків з іншими об'єктами тощо. Метою роботи є створення такого набору даних. Для цього сформульовано завдання:

1) скомпілювати набір даних з різних джерел;

2) розробити протокол перевірки релевантності інформації; ця перевірка покликана оцінити якість отриманого набору даних;

3) виконати перевірку відповідно до розробленого протоколу.

Отже, подальший розвиток класифікації зображень завдяки мультимодальному аналізу потребує створення мультимодального набору даних, що використовуватиме як самі зображення, так і текстову інформацію про зображення, що дасть змогу застосовувати останні віхи розвитку аналізу природної мови за допомогою мовних нейронних мереж для виконання завдання класифікації.

Серед публічно доступних наборів даних для навчання нейронний мереж уже існують деякі схожі набори, але жоден з них не оптимізований під описане вище завдання. Кожен з наведених наборів має переваги й недоліки для завдання класифікації зображень.

Набір *WIT* (*Wikipedia-based Image Text*) [18] є великою множиною даних і містить таку інформацію:

– зображення;
– мова статті Вікіпедії;
– заголовок статті;
– назва секції, де розміщене зображення;
– підпис зображення, зокрема й альтернативний текст, що відтворюється на Web-сторінці внаслідок наведення курсором на зображення й може відрізнятись від основного підпису;
– розміри зображення та його тип даних (*MIME-type*);
– контекстний опис зі сторінки;
– контекстний опис із секції;
– інша метаінформація.

У цьому наборі даних контекстні описи зображень наводяться таким чином, щоб текст максимально точно описував саме зображення. Це зроблено завдяки ручному вибору необхідних елементів сторінки Вікіпедії та секції сторінки в такий спосіб, щоб обраний текст описував зображення.

Важливою особливістю набору даних *WIT* є його розмір, а саме близько 3,5 млн елементів. Це дає змогу, відповідно до заяви авторів, використовувати цей набір даних для тренування мультимодальних моделей з нуля, тобто кількість елементів дає змогу навчати моделі загальних мультимодальних ознак, на основі яких можна добудовувати моделі під конкретні дані та завдання методом донавчання.

Недоліком набору даних *WIT* у розв'язанні завдання класифікації зображень є відсутність групування елементів за якісно виокремленими класами. На відміну від наборів даних, створених

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2025. No. 1 (31)

спеціально для завдання класифікації зображень, *WIT* має безліч окремих зображень, а отже, не допомагає будувати моделі класифікації. Виокремлення класів із галузей, яким присвячені сторінки Вікіпедії, не є можливим з двох причин.

1. Галузі, описані на сторінках Вікіпедії, доволі загальні й не є конкретними класами об'єктів зображень, а скоріше є загальними галузями знань людства.

2. Для кожного конкретного класу зображень у наборі даних міститься незначна кількість елементів, недостатня для того, щоб навчити модель ні за умови підходу навчання з нуля, ні в разі донавчання моделі для конкретного завдання.

Загалом набір даних *WIT* не ефективний у розв'язанні завдання класифікації зображень, проте застосовує підходи до збирання інформації, що може бути корисною для цієї проблеми.

Набір даних *Flickr30K* [19] має меншу кількість інформації, але їй властиві певні переваги. Цей набір даних містить таку інформацію:

– зображення;

– декілька (в середньому близько п'яти) стислих описів цього зображення.

Описи зображень зазначеного набору даних створені операторами-людьми, тому точно описують зображення, без помилок, пов'язаних з ідентифікацією об'єктів чи дій на зображенні. Набір даних *Flickr30K* містить 31 783 зображення та 158 915 описів.

Flickr30K так само, як набір *WIT,* не має конкретних класів зображень, однак через особливість інформації, унаслідок використання наданих текстових описів із цього набору можна виокремити певну кількість класів зображень, оскільки зображення цього набору є менш різноманітними, ніж у *WIT*.

Отже, серед наявних наборів даних є такі, що можуть бути застосовані в навчанні моделі нейронної мережі для завдання класифікації зображень. Проте ці набори не є готовими до такого навчання й або потребують складного попереднього оброблення, або можуть надавати загальні ідеї щодо збору більш підхожих наборів даних.

Матеріали й методи

Методи й матеріали дослідження варто розмежувати відповідно до його етапів, а саме: збирання набору даних і його оброблення та оцінювання релевантності набору даних щодо класів зображень цього набору.

Для отримання набору даних необхідно взяти джерело зображень, джерело текстових описів зображень та джерело загальних текстових описів класів зображень окремо від конкретних прикладів.

Джерелом зображень був обраний набір даних *ImageNet*. Завдяки великому розміру, значній кількості класів зображень (1000 класів) та широкому застосуванню в розв'язанні завдання класифікації зображень цей набір даних допомагає легко добудовувати необхідну додаткову інформацію, має широку підтримку реалізацій у бібліотеках для навчання нейронних мереж та дає змогу порівнювати виконане завдання із наявними результатами завдяки активній спільноті дослідників, які застосовують *ImageNet* як міру якості роботи моделей класифікації [20].

ImageNet має доволі просту структуру, у якій для кожного елемента набору є така інформація:

– зображення;

– клас зображення;

– інші метадані.

Загалом *ImageNet* покликаний розв'язувати не лише завдання класифікації. Але оскільки це дослідження зосереджене саме на цьому завданні, тому опис всіх можливостей набору даних *ImageNet* не наводиться.

За основу інформації текстових описів зображень обрано набір даних *ImageNet-Captions* [21]. Він створений на основі *ImageNet* і є набором текстових анотацій до зображень, взятих із сайту розміщення фотографій *Flickr*. Набір містить окремо назви зображень та їх описи. Для простоти роботи з ним вся інформація, назви і описи сприймаються як одне текстове поле, тобто стрічки конкатенуються. Сайт *Flickr* є джерелом зображень для набору даних *Flickr30K*, згаданого вище, а також для набору даних *ImageNet*. Тестові анотації є описами зображень на сайті, які створили автори зображень під час завантаження на сайт. Цей підхід дав змогу авторам *ImageNet-Captions* розробити мультимодальний набір даних без застосування додаткових ресурсів з анотування зображень людьми-операторами, що є складним і недешевим процесом.

Недоліком анотацій *ImageNet-Captions* є нижча якість текстової інформації. Оскільки дані не були згенеровані спеціально для опису зображень, а скоріше є текстовими частинами дописів авторів на сайті, що супроводжують відповідні зображення, ці анотації не завжди є релевантними для навчання

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 1 (31) ISSN 2524-2296 (online)

нейронних мереж, адже можуть не додавати інформацію, яка стосується самих зображень, а лише створювати шум. Розв'язання проблеми релевантності текстових даних для зображень описано далі.

Джерелом текстової інформації для опису окремих класів взяті фрагменти сторінок Вікіпедії англійською мовою, що описують клас об'єкта на зображенні. Наприклад, для опису класу *lesser_panda* взяті фрагменти зі сторінки Вікіпедії під назвою *Red panda*. Пошук сторінок для цитування виконувався автоматично, де це можливо. Більшість класів *ImageNet* мають таку назву, яка напряму відповідає одній конкретній сторінці Вікіпедії. Якщо ж назва вимагала вибору з декількох сторінок чи пошук потрапляв на переадресацію сторінки або ж не знаходив потрібної сторінки взагалі, вона обиралася в ручному режимі.

З кожної сторінки для набору даних використано секцію підсумкового опису (*summary*). Також на сторінках, що мали спеціальну секцію з описом об'єкта, а саме з назвою *Description*, що властиво сторінкам, де описано види чи породи тварин, вміст цієї секції також був доданий до текстового опису класу. У такий спосіб для всієї тисячі класів було зібрано якісні описи об'єктів, що корелюють із інформацією, яку може вивчити нейронна мережа, що обробляє зображення, а також із загальною інформацією про об'єкти, їх властивості та взаємозв'язки.

Комбінація гарантовано релевантних описів класів та дещо релевантних описів зображень дає змогу проаналізувати релевантність текстових описів для підвищення ефективності навчання моделей на зібраному наборі даних.

Для оцінювання релевантності текстових описів зображень розроблений метод, що застосовує гарантовано релевантні описи класів зображень. Зазначений метод, оснований на окремій моделі машинного навчання, має на меті відфільтрувати менш релевантні текстові приклади набору даних. Алгоритм оцінювання релевантності передбачає певні кроки.

1. Створити набір, що містить класи зображень, текстові описи зображень та випадковим чином обрані речення з описів класів за принципом одне речення на один елемент набору даних.

2. На основі навченої мовної моделі донавчити модель, яка буде отримувати на вхід сконкатеновані опис зображення та речення з опису класу. Модель на вході віддає номер класу зображення.

3. За допомогою навченої моделі для кожного опису зображення, що є в наборі даних *ImageNet-Captions*, оцінити релевантність опису до класу зображення. Вилучити описи, що показують низьку релевантність з кінцевого набору даних. Для тих описів, що залишились, зберегти коефіцієнт впевненості моделі в релевантності опису до класу.

Отже, послідовно розглянемо кроки алгоритму, їх особливості та визначимо, які методи й матеріали необхідні для виконання кожного з них.

Створення тренувальної та тестової вибірки для навчання моделі-предиктора релевантності полягає у виділенні певної кількості інформації на кожну вибірку з набору даних *ImageNet-Captions*. Весь набір містить 445 984 зображення, для яких є або назва, або опис, або і те, та інше. Крім цього, зі всіх елементів 389 598 мають назву або опис, що містить більш ніж одне слово. Саме ці елементи взяті за основу набору даних для тренування моделі-предиктора релевантності. Ці елементи розділені на 999 класів набору *ImageNet*. Розподіл даних за класами приблизно відповідає нормальному розподілу з незначними відхиленнями (див. рис. 1). На рисунку зображена відповідність кількості прикладів для одного класу (вісь *X*) до кількості класів із такою кількістю прикладів (вісь *Y*).

Для тренувальної вибірки було випадковим чином обрано рівно 300 тис. елементів. Решта 89 598 елементів, тобто приблизно 23 %, становили тестову збірку. У цьому разі для покращення роботи моделі на менш релевантних класах, що містять 100 прикладів чи менше, гарантовано мінімум половина прикладів опинилась у тренувальній збірці, і мінімум 20 % прикладів опинились у тестовій.

Наступним кроком у створенні набору даних для тренування моделі-предиктора релевантності є виокремлення речень з описів класів зображень, зібраних із сторінок Вікіпедії. Для кожного класу зображень було виокремлено від двох до десяти речень. За наявності обирались найдовші речення із зібраних текстів. Далі ці речення були додані до текстових описів зображень. Речення з описів класів додані в кінець опису зображень. У цьому разі кожне речення, поєднане з кожним описом зображення, тобто множина отриманих текстових елементів набору даних, – це декартів добуток множини речень і множини описів зображень. Приклади елементів набору даних наведені в табл. 1. Опис класів у тексті скорочений заради простоти читання в цій статті.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online)
Innovative technologies and scientific solutions for industries. 2025. No. 1 (31)

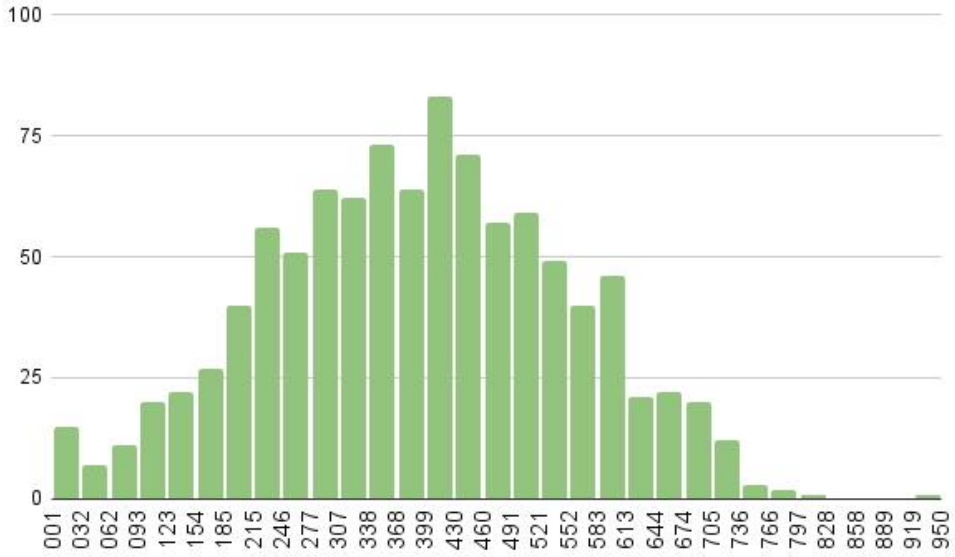

Рис. 1. Розподіл кількостей елементів набору даних *ImageNet-Captions* за класами

Таблиця 1. *Приклади елементів набору даних для тренування моделі-предиктора релевантності*

Клас *ImageNet* (wnid / назва)	**Текст**
n07693725 / bagel	Homemade bagels-pre boiling I made the dough yesterday and let them rise in the 'fridge overnight, as per the recipe. i just pulled these out of the fridge and realize that i made them WAY too big. A bagel (Yiddish: בייגל, romanized: beygl; Polish: bajgiel [ˈbajgjɛl] ; also spelled beigel) is a bread roll originating in the Jewish communities of Poland. Bagels are traditionally made from yeasted wheat dough that is shaped by hand into a torus or ring, briefly boiled in water, and then baked.
n04462240 / toyshop	In the toy-shop Barbie, born 1959, still going strong A toy store or toy shop is a type of retail business specializing in selling toys.
n02277742 / ringlet	Ringlet (Aphantopus hyperantus) Ringlets are quite abundant at the moment - and seem to be a lot more approachable than most butterflies I've been attempting to photograph. Handheld, natural light The ringlet (Aphantopus hyperantus) is a butterfly in the family Nymphalidae. It is only one of the numerous 'ringlet' butterflies in the tribe Satyrini.

Зрештою було отримано набір даних, що містить близько 1,7 млн елементів, 1,3 млн в тренувальній вибірці і 400 тис. у тестовій вибірці.

Для створення моделі-предиктора релевантності обрано модель *RoBERTa*, попередньо натреновану на великих корпусах даних. Модель приймає на вхід текст та видає на виході векторне подання цього тексту. Векторне подання має форму послідовності ембедингів (тобто векторів) за принципом: один вектор на одне вхідне слово (токен). Класичним методом отримання ембедингу тексту з ембедингів окремих слів тексту є пулінг методом середнього, тобто з усіх векторів тексту береться поелементне середнє значення. Існують і більш складні стратегії пулінгу. Проте дослідження в галузі мультимодального оброблення текстової та графічної інформації показують, що коректно підібрана проста стратегія, така як пулінг методом середнього, може давати вищу якість кінцевої моделі, на відміну від більш складних стратегій [22].

Формула (1) демонструє принцип роботи пулінгу методом середнього.

$$x_i = \sum_{n=1}^{N} a_{in} \Big/ N , \qquad (1)$$

де x_i – i -й елемент ембедингу тексту; a_{in} – i -й елемент ембедингу n -го слова; N – кількість токенів.

Для того щоб перетворити ембединг тексту в клас зображення, використовується щільний шар нейронів, який генерує вірогідності того, що цей ембединг є конкретним класом. Тобто щільний шар генерує вектор розмірністю 999, де кожне значення – це вірогідність того, що вхідний текст описує саме клас із цим порядковим номером. Для отримання нормалізованих вірогідностей значення кожного елемента отриманого вектора обробляється функцією сигмоїди, перетворюючи їх на значення в діапазоні

Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 1 (31)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

від 0 до 1. Для контролю перенавчання на певну характеристику ембедингу між операцією пулінгу та щільним шаром використана операція викидання (*dropout*) з вірогідністю викидання кожного конкретного значення $P_{dropout} = 0,25$. На рис. 2 зображена високорівнева архітектура моделі.

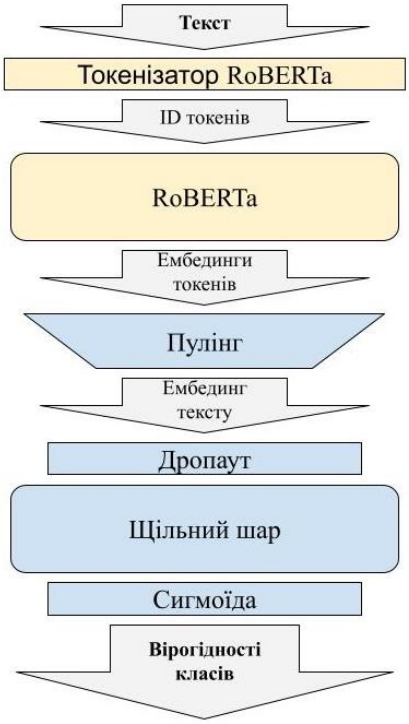

Рис. 2. Високорівнева архітектура моделі-предиктора релевантності

На рисунку напівжирним виділені вхідні та вихідні дані. Жовтим кольором позначені попередньо навчені елементи, а блакитним – додані елементи.

Запропонована модель була навчена на тренувальній вибірці. Тестова вибірка використовувалась для контролю перенавчання на всіх епохах навчання. Кожна епоха охоплювала 75 % тренувального набору, обраного та відсортованого випадковим чином. Безперервне тренування зайняло 120 епох і близько 80 год.

Унаслідок тренування отримано модель, що приймає на вхід текстову інформацію, а на вихід видає вірогідності того, що цей текст описує конкретний клас з набору даних *ImageNet*. Завдяки цій моделі можна оцінити релевантність кожного окремого опису зображення для кожного класу. Для цього необхідне виконання певних дій.

1. Обробити кожен текстовий опис зображення за допомогою отриманої моделі без додавання текстової інформації про сам клас зображення.

2. Для кожного обробленого прикладу взяти значення вірогідності того, що текст належить до того класу, до якого він насправді належить $P_{correct}$.

3. Якщо вірогідність $P_{correct}$ нижча за середнє значення вірогідностей класів для цього прикладу, текстова інформація відкидається.

4. Якщо вірогідність $P_{correct}$ нижча за 0,5, текстова інформація відкидається.

5. Якщо вірогідність $P_{correct}$ перевищує чи дорівнює 0,5, текстова інформація зберігається в елементі набору даних як коефіцієнт релевантності тексту. Коефіцієнти релевантності тексту можуть використовуватись надалі в навчанні нейронних мереж чи в інших методах, де застосовується отриманий набір даних.

6. Для покращення результатів подальшого використання коефіцієнтів релевантності впроваджується нормалізація. Обране порогове значення $P_{correct} = 0,9$. Якщо значення релевантності тексту вище, ніж це порогове значення, то йому присвоюється значення 1. До інших значень додається 0,1. Потім усі значення нормалізуються таким чином, щоб вони були рівномірно розподілені на відрізку $(0; 1]$.

Отже, у дослідженні розроблено метод оцінювання релевантності текстових описів зображень для завдань класифікації на основі мультимодального набору даних. Джерелом зображень обрано *ImageNet*, текстових описів – *ImageNet-Captions* та Вікіпедія. Для підвищення якості інформації застосовано спеціалізовані методи оцінювання текстів, основані на навчанні мовної моделі *RoBERTa* та на різниці релевантностей різних джерел текстової інформації.

Результати досліджень та їх обговорення

Дослідження передбачало два аспекти розвитку окресленого питання, а саме: розроблення методу покращення якості текстової інформації способом відсіювання шуму, тобто менш релевантних даних, які сповільнюють навчання нейронних мереж і пригнічують якість статистичних та інших конвенційних моделей; створення набору даних, що комбінує текстову метаінформацію про об’єкти на зображеннях із самими зображеннями.

Можливість навчання такої моделі основана на наявності більш релевантних описів класів, що допомагають тимчасово підвищити релевантність

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2025. No. 1 (31)

описів зображень, надаючи цим нейронній мережі необхідний сигнал. Завдяки наявності гарантовано релевантних даних модель може вловлювати деталі описів зображень. Крім цього, якщо додане до опису зображення речення є ключовим у визначенні класу зображення, а внаслідок вилучення цього речення опис зображення перестає бути корисним для визначення класу, то опис зображення можна вважати нерелевантним і відкинути.

Після навчання моделі-предиктора релевантності й застосування її на даних набору *ImageNet-Captions* отримано набір даних, що містить 28 610 елементів. Завдяки цьому підходу вдалось вилучити з набору найбільш шумні, тобто найменш релевантні елементи, такі як технічні описи фотоапаратів, що використовувались для створення зображень, гештеги та інша метаінформація, не пов'язана із зображеними об'єктами.

Коефіцієнти релевантності текстових анотацій до зображень є цінним артефактом розробленого алгоритму створення набору даних. Ці значення, що рівномірно розподілені між 0 і 1, позначають рівень того, наскільки текст пов'язаний із зображенням. У процесі тренування класифікатора ця інформація буде корисною для модифікації функції помилки навчання. Отже, функція помилки зможе брати до уваги не лише різницю між правдивим значенням цільової змінної та передбаченим значенням, а також і те, наскільки можна очікувати, що вхідна інформація якісна. Це має сприяти швидшому навчанню моделі, адже приклади з вищим рівнем шуму менше впливатимуть на перебіг навчання, ніж більш релевантні приклади. Такий підхід також дає змогу підвищити ефективність запропонованої моделі.

Висновки й перспективи подальшого розвитку

Набори даних, що комбінують у собі два чи більше видів інформації, здатні надавати ширший контекст для розв'язання завдань, що, як правило, асоціюються лише з одним типом даних. Це сприяє більш ефективному впровадженню методів машинного навчання. У межах цієї роботи створено новий набір даних, на основі якого можуть проводитись подальші дослідження мультимодальних моделей.

Розроблений метод фільтрування текстових даних дає змогу завдяки публічно доступній інформації, без додаткового ручного анотування зібрати текстові метадані для зображень. Завдяки цьому методу відкриваються нові можливості для створення більшого за розміром та обсягом покриття, кращих за якістю різноманітних характеристик зображених об'єктів наборів текстових і графічних даних.

Потенціал подальшого покращення отриманого набору даних полягає в більш оптимальному підборі гіперпараметрів алгоритму, а саме порогових значень вірогідності релевантності, як нижнього (нижче якого анотації вважаються нерелевантними), так і верхнього (вище якого анотації вважаються абсолютно релевантними).

Також зміна розподілу коефіцієнтів релевантності з лінійної на більш складну криву може привести до ефективнішого корегування релевантності.

Подальшого розвитку також потребують методи автоматизованого збору даних без участі людей-операторів для анотації, більш оптимізовані методи вибору релевантної текстової інформації для окремих класів зображень та аналогічний метод фільтрування нерелевантних анотацій зображень без прив'язки до конкретного класу.

Отже, запропонований метод дасть змогу розширити обсяги публічно доступних наборів даних, що містять зображення та тексти. Кінцевою метою цього напряму дослідження було б повністю проанотоване створення набору даних із зображеннями, наближеним за своїм розміром до *ImageNet*, а за якістю анотацій – до *Flickr30K*. Мета роботи – створення мультимодального набору даних для класифікації зображень – досягнута.

Список літератури

1. Mensink T., Verbeek J., Perronnin F., Csurka G. Distance-Based image classification: generalizing to new classes at near-zero cost. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2013. № 35 (11). P. 2624–2637. DOI: https://doi.org/10.1109/tpami.2013.83
2. Xu Z., Sun K., Mao J. Research on ResNet101 network chemical reagent label image classification based on transfer learning. *IEEE Xplore*. 2020. DOI: https://doi.org/10.1109/ICCASIT50869.2020.9368658

Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 1 (31)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

3. Tang X., Zhou C., Chen L., Wen Y. Enhancing medical image classification via augmentation-based pre-training. *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2021. DOI: https://doi.org/10.1109/bibm52615.2021.9669817
4. Dao H.N., Nguyen T., Cherubin Mugisha, Paik I. A multimodal transfer learning approach using pubmedclip for medical image classification. *IEEE Access*. 2024. №12. P. 75496–75507. DOI: https://doi.org/10.1109/access.2024.3401777
5. Ma M., Ma W., Jiao L., Liu X., Liu F., Li L., Yang S. MBSI-Net: multimodal balanced self-learning interaction network for image classification. *IEEE Transactions on Circuits and Systems for Video Technology*. 2024. №34(5). P. 3819–3833. DOI: https://doi.org/10.1109/tcsvt.2023.3322470
6. Chen Q., Shi Z., Zuo Z., Fu J., Sun Y. Two-Stream hybrid attention network for multimodal classification. *IEEE International Conference on Image Processing (ICIP)*. 2021. DOI: https://doi.org/10.1109/icip42928.2021.9506177
7. "ImageNet". URL: www.image-net.org (дата звернення: 10.10.2024).
8. Liu Y., Ott M., Goyal N., Du J., Joshi M.S., Chen D., Levy O., Lewis M., Zettlemoyer L., Stoyanov V. RoBERTa: A robustly optimized bert pretraining approach. *arXiv (Cornell University)*. 2019. DOI: https://doi.org/10.48550/arxiv.1907.11692
9. Satheesh Kumar NJ, CH A. DRCNN-WS: a novel approach for high-resolution video using recurrent neural networks and walrus search. *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*. 2024. P. 1–6. DOI: https://doi.org/10.1109/icrito61523.2024.10522118
10. Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. An image is worth 16x16 words: transformers for image recognition at scale. *arXiv.org*. 2021. DOI: https://doi.org/10.48550/arXiv.2010.11929
11. Arpit Bansal Mathematics, Kumar K., Singla S. Multimodal deep learning: integrating text and image embeddings with attention mechanism. *IEEE Xplore*. 2024. DOI: https://doi.org/10.1109/aiiot58432.2024.10574665
12. Radford A., Kim J.W., Hallacy C., Ramesh A., Goh G., Agarwal S., Sastry G., Askell A., Mishkin P., Clark J., Krueger G., Sutskever I. Learning transferable visual models from natural language supervision. *arXiv.org*. 2021. DOI: https://doi.org/10.48550/arXiv.2103.00020
13. Peng L., Jian S., Li D., Shen S. MRML: Multimodal rumor detection by deep metric learning. *IEEE Xplore*. 2023. DOI: https://doi.org/10.1109/ICASSP49357.2023.10096188
14. Guo W., Wang J., Wang S. Deep multimodal representation learning: a survey. *IEEE Access*. 2019. №7. P. 63373–63394. DOI: https://doi.org/10.1109/access.2019.2916887
15. Karpathy A., Fei-Fei L. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2017. №39(4). P. 664–676. DOI: https://doi.org/10.1109/tpami.2016.2598339
16. Shin S., Jang J., Jung M., Kim J., Jung Y., Jung H. Construction of a machine learning dataset for multiple AI tasks using korean commercial multimodal video clips. *ICTC*. 2020. P.1264–1266. DOI: https://doi.org/10.1109/ictc49870.2020.9289319
17. Chen B., Liu J., Li Z., Yang M. Seeking the sufficiency and necessity causal features in multimodal representation learning. *arXiv (Cornell University)*. 2024. DOI: https://doi.org/10.48550/arxiv.2408.16577
18. Srinivasan K., Raman K., Chen J., Bendersky M., Najork M. WIT: wikipedia-based image text dataset for multimodal multilingual machine learning. *SIGIR '21: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021. P. 2443–2449. DOI: https://doi.org/10.1145/3404835.3463257
19. Young P., Lai A., Hodosh M., Hockenmaier, J. From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*. 2014. №2. P. 67–78. DOI: https://doi.org/10.1162/tacl_a_00166
20. "Papers with Code – ImageNet Benchmark (Image Classification)". URL: https://paperswithcode.com/sota/image-classification-on-imagenet (дата звернення: 01.11.2024).
21. Fang A., Ilharco G., Wortsman M., Wan Y., Shankar V., Dave A., Schmidt L. Data determines distributional robustness in contrastive language image pre-training (CLIP). *arXiv (Cornell University)*. 2022. DOI: https://doi.org/10.48550/arxiv.2205.01397
22. Chen J., Hu H., Wu H., Jiang Y., Wang C. Learning the best pooling strategy for visual semantic embedding. *arXiv (Cornell University)*. 2020. DOI: https://doi.org/10.48550/arxiv.2011.04305

References

1. Mensink, T., Verbeek, J., Perronnin, F., Csurka, G. (2013), "Distance-Based image classification: generalizing to new classes at near-zero cost". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, № 35(11), P. 2624–2637. DOI: https://doi.org/10.1109/tpami.2013.83

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2025. No. 1 (31)

2. Xu, Z., Sun, K., Mao, J. (2020), "Research on ResNet101 Network chemical reagent label image classification based on transfer learning". *IEEE Xplore.* DOI: https://doi.org/10.1109/ICCASIT50869.2020.9368658
3. Tang, X., Zhou, C., Chen, L., Wen, Y. (2021), "Enhancing medical image classification via augmentation-based pre-training". *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).* DOI: https://doi.org/10.1109/bibm52615.2021.9669817
4. Dao, H.N., Nguyen, T., Cherubin Mugisha, Paik, I. (2024), "A multimodal transfer learning approach using pubmedclip for medical image classification". *IEEE Access*, 12, P. 75496–75507. DOI: https://doi.org/10.1109/access.2024.3401777
5. Ma, M., Ma, W., Jiao, L., Liu, X., Liu, F., Li, L., Yang, S. (2024), "MBSI-Net: multimodal balanced self-learning interaction network for image classification". *IEEE Transactions on Circuits and Systems for Video Technology*, № 34(5), P. 3819–3833. DOI: https://doi.org/10.1109/tcsvt.2023.3322470
6. Chen, Q., Shi, Z., Zuo, Z., Fu, J., Sun, Y. (2021), "Two-stream hybrid attention network for multimodal classification". *2022 IEEE International Conference on Image Processing (ICIP)*. DOI: https://doi.org/10.1109/icip42928.2021.9506177
7. "ImageNet". available at: https://www.image-net.org/
8. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M.S., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019), "RoBERTa: a robustly optimized bert pretraining approach". *arXiv (Cornell University)*, № 1. DOI: https://doi.org/10.48550/arxiv.1907.11692
9. Satheesh Kumar NJ, CH, A. (2024), "DRCNN-WS: A novel approach for high-resolution video using recurrent neural networks and walrus search". *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, P. 1–6. DOI: https://doi.org/10.1109/icrito61523.2024.10522118
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N. (2021), "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". *arXiv.org.* DOI: https://doi.org/10.48550/arXiv.2010.11929
11. Arpit Bansal Mathematics, Kumar, K., Singla, S. (2024), "Multimodal deep learning: integrating text and image embeddings with attention mechanism". *IEEE Xplore.* DOI: https://doi.org/10.1109/aiiot58432.2024.10574665
12. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I. (2021), "Learning transferable visual models from natural language supervision". *arXiv.org.* DOI: https://doi.org/10.48550/arXiv.2103.00020
13. Peng, L., Jian, S., Li, D. and Shen, S. (2023), "MRML: multimodal rumor detection by deep metric learning". *IEEE Xplore.* DOI: https://doi.org/10.1109/ICASSP49357.2023.10096188
14. Guo, W., Wang, J., Wang, S. (2019), "Deep multimodal representation learning: a survey". *IEEE Access*, № 7, P. 63373–63394. DOI: https://doi.org/10.1109/access.2019.2916887
15. Karpathy, A., Fei-Fei, L. (2017), "Deep visual-semantic alignments for generating image descriptions". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, № 39 (4), P. 664–676. DOI: https://doi.org/10.1109/tpami.2016.2598339
16. Shin, S., Jang, J., Jung, M., Kim, J., Jung, Y., Jung, H. (2020), "Construction of a machine learning dataset for multiple AI tasks using korean commercial multimodal video clips". *ICTC*. P. 1264–1266. DOI: https://doi.org/10.1109/ictc49870.2020.9289319
17. Chen, B., Liu, J., Li, Z., Yang, M. (2024), "Seeking the Sufficiency and Necessity Causal Features in Multimodal Representation Learning". *arXiv (Cornell University).* DOI: https://doi.org/10.48550/arxiv.2408.16577
18. Srinivasan, K., Raman, K., Chen, J., Bendersky, M., Najork, M. (2021), "WIT: wikipedia-based image text dataset for multimodal multilingual machine learning". *SIGIR '21: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 2021. P. 2443–2449.DOI: https://doi.org/10.1145/3404835.3463257
19. Young, P., Lai, A., Hodosh, M., Hockenmaier, J. (2014), "From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions". *Transactions of the association for computational linguistics*, № 2, P. 67–78. DOI: https://doi.org/10.1162/tacl_a_00166
20. "Papers with Code - ImageNet Benchmark (Image Classification)". available at: https://paperswithcode.com/sota/image-classification-on-imagenet
21. Fang, A., Ilharco, G., Wortsman, M., Wan, Y., Shankar, V., Dave, A., Schmidt, L. (2022), "Data Determines Distributional Robustness in Contrastive Language Image Pre-training (CLIP)". *arXiv (Cornell University).* DOI: https://doi.org/10.48550/arxiv.2205.01397
22. Chen, J., Hu, H., Wu, H., Jiang, Y., Wang, C. (2020), "Learning the Best Pooling Strategy for Visual Semantic Embedding". *arXiv (Cornell University).* DOI: https://doi.org/10.48550/arxiv.2011.04305

Надійшла (Received) 02.01.2025

Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 1 (31)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

Відомості про авторів / About the Authors

Дашенков Дмитро Сергійович – Харківський національний університет радіоелектроніки, аспірант кафедри програмної інженерії, Харків, Україна; e-mail: dmytro.dashenkov@nure.ua; ORCID ID: https://orcid.org/0000-0001-9797-1863

Смеляков Кирило Сергійович – доктор технічних наук, Харківський національний університет радіоелектроніки, професор кафедри програмної інженерії, Харків, Україна; e-mail: kyrylo.smelyakov@nure.ua; ORCID ID: https://orcid.org/ 0000-0001-9938-5489

Dashenkov Dmytro – Kharkiv National University of Radio Electronics, PhD Student at the Software Engineering department, Kharkiv, Ukraine.

Smelyakov Kirill – Doctor of Sciences (Engineering), Kharkiv National University of Radio Electronics, Professor at the Software Engineering department, Kharkiv, Ukraine.

EXTENDING THE IMAGENET DATASET FOR MULTIMODAL TEXT AND IMAGE LEARNING

Subject matter: image processing methods for classification and other computer vision tasks using multimodal data, including text descriptions of classes and images. **Goal:** development of a multimodal dataset for image classification using textual meta-information analysis. The resulting dataset should consist of image data, image classes, namely 1000 classes of objects depicted in photos from the ImageNet set, textual descriptions of individual images, and textual descriptions of image classes as a whole. **Tasks:** 1) based on the images of the ImageNet dataset, compile a dataset for training classifier models with text descriptions of image classes and individual images; 2) based on the obtained dataset, conduct an experiment on training a language neural network to confirm the effectiveness of using this approach to solve the classification problem. **Methods:** compilation of datasets manually, training of speech neural networks based on the RoBERTa architecture. The neural network training was carried out using the fine-tuning method, namely, adding a neural network layer to an existing model to obtain a new machine learning model capable of performing the selected task. **Results:** the result of the work is the creation of a dataset that combines image data with text data. The resulting dataset is useful for establishing a connection between the information that a machine learning model is able to extract from photos and the information that the model can extract from text data. The multimodal approach can be used to solve a wide range of problems, as demonstrated by the example of training a language neural network. The trained language model processes the descriptions of images contained in the dataset and predicts the class of the image to which this description is associated. The model is designed to filter out irrelevant text metadata, improving the quality of the dataset. **Conclusions:** data sets that combine multiple types of data can provide a broader context for solving problems that are typically associated with only one type of data, allowing for more effective application of machine learning methods.

Keywords: multimodal machine learning; image classification; natural language processing; datasets; text metadata.

Бібліографічні описи / Bibliographic descriptions

Дашенков Д. С., Смеляков К. С. Розширення набору даних *ImageNET* для мультимодального навчання з текстом та зображеннями. *Сучасний стан наукових досліджень та технологій в промисловості*. 2025. № 1 (31). С. 20–31. DOI: https://doi.org/10.30837/2522-9818.2025.1.020

Dashenkov, D., Smelyakov, K (2025), "Extending the *ImageNET* dataset for multimodal text and image learning", *Innovative Technologies and Scientific Solutions for Industries*, No. 1 (31), P. 20–31. DOI: https://doi.org/10.30837/2522-9818.2025.1.020