DOI: 10.58254/viti.8.2025.11.133

УДК 004.8:004.021.64:519.85

Меркотан Д. Ю. ORCID: 0000-0003-1425-9948 (ВІТІ ім. Героїв Крут)
канд. техн. наук, доцент Троцько О. О. ORCID: 0000-0001-7535-5023 (ВІТІ ім. Героїв Крут)

МЕТОД ГРАФОВОГО ЗЛИТТЯ МУЛЬТИМОДАЛЬНИХ ДАНИХ З ВИКОРИСТАННЯМ БАГАТОЗАДАЧНОГО НАВЧАННЯ

У реальних умовах дані зазвичай містять декілька модальностей та можуть мати неексклюзивні мітки. Ключовим етапом мультимодального навчання є процес злиття мультимодальної інформації, оскільки він забезпечує об’єднання ознак з різних джерел у спільний векторний простір. Це дає змогу класифікатору використовувати сформований інтегрований вектор для отримання кінцевого прогнозу. Водночас традиційні методи мультимодального злиття рідко беруть до уваги міжмодальні взаємодії, які відіграють важливу роль у виявленні залежностей між модальностями та побудові єдиного простору їх інтеграційного представлення.

У цій роботі запропоновано метод графового злиття мультимодальних даних з використанням багатозадачного навчання. Він спрямований на формування спільного простору інтеграційного представлення для всіх міжмодальних взаємодій та на адаптивне налаштування функцій втрат окремих завдань з метою досягнення оптимальних показників ефективності обробки мультимодальних даних. Розроблений метод використовує вдосконалену графову мережу мультимодального злиття, яка враховує міжмодальні взаємодії між усіма комбінаціями модальностей та динамічно розподіляє вагові коефіцієнти для кожної пари модальностей залежно від конкретного зразка даних. Крім того, впроваджено новий підхід багатозадачного навчання для розв’язання проблем багатомітковості шляхом автоматичного регулювання процесу навчання, як на рівні завдань, так і на рівні окремих зразків.

Експериментальні результати засвідчують, що запропонований метод перевищує ефективність базових моделей та окремих сучасних методів. Також продемонстровано гнучкість і модульність запропонованих компонентів мультимодального злиття та динамічного багатозадачного навчання, що дозволяє інтегрувати їх у різні типи нейронних мереж.

Ключові слова: *машинне навчання, злиття даних, мультимодальне злиття, обробка даних, багатозадачне навчання, штучний інтелект, нейронні мережі, інформаційні системи.*

D. Merkotan, O. Trotsko. Method of graph fusion of multimodal data using multitask learning

In real-world conditions, data typically contain multiple modalities and may have non-exclusive labels. A key stage of multimodal learning is the process of multimodal fusion, as it enables the integration of features from different sources into a unified vector space. This allows the classifier to utilize the constructed integrated vector to produce the final prediction. At the same time, traditional multimodal fusion methods rarely take into account cross-modal interactions, which play an essential role in uncovering dependencies between modalities and in constructing a shared space of their integrated representation.

In this paper, we propose a conceptual framework for multimodal fusion with the use of multi-task learning. It is aimed at modeling a joint integrated representation space for all cross-modal interactions and adaptively tuning the loss functions of individual tasks in order to achieve optimal performance. The developed model employs a novel hierarchical multimodal fusion network that captures cross-modal interactions across all modality combinations and dynamically allocates weight coefficients for each pair depending on the specific data sample.

In addition, a new multi-task learning approach is introduced to address multi-label classification challenges by automatically adjusting the training process both at the task level and at the sample level. Experimental results demonstrate that the proposed conceptual framework outperforms baseline models as well as several state-of-the-art methods. Furthermore, the flexibility and modularity of the proposed components of multimodal fusion and dynamic multi-task learning are showcased, making them applicable to various types of neural network architectures.

Keywords: *machine learning, data fusion, multimodal fusion, data processing, multi-task learning, artificial intelligence, neural networks, information systems.*

Актуальність та постановка завдання в загальному вигляді. Мультимодальне навчання привертає значну увагу наукової спільноти завдяки своїй здатності ефективно використовувати великі обсяги реальних даних, що зазвичай містять декілька джерел інформації [1–6]. На відміну від підходів з використанням лише однієї модальності, мультимодальне навчання орієнтоване на розкриття багатого інформаційного потенціалу, що міститься у різних вхідних модальностях.

Одним із ключових етапів мультимодального навчання є злиття даних із різних модальностей, у процесі якого вхідні ознаки кожної модальності об’єднуються для формування єдиного векторного представлення. Відтак спосіб, у який здійснюється злиття ознак, суттєво впливає на здатність моделі ефективно засвоювати інформацію, надану кількома джерелами вхідних даних.

На відміну від традиційного уявлення, просте збільшення кількості вхідних модальностей не завжди приводить до кращих результатів [5]. Основною причиною зниження ефективності є ігнорування міжмодальних взаємодій.

Ефективне злиття представлень різнорідних модальностей перетворилося на актуальну проблему мультимодального навчання та, відповідно, привернуло значну увагу наукової спільноти. Гетерогенна природа мультимодальних даних створює суттєвий бар’єр на шляху до використання повної сукупності інформації з усіх модальностей, що є ключовим для глибокого розуміння та ефективного застосування насичених мультимедійних даних [7].

Перші спроби мультимодального злиття ґрунтувалися на окремому опрацюванні кожної модальності. Кожна модальність опрацьовувалася у власній мережі, а проміжні ознаки поєднувалися на різних етапах обробки – зокрема у формах раннього злиття та пізнього злиття [8]. Однак через гетерогенність мультимодальних даних та відсутність узгодженості між мережами, отримане інтегроване векторне представлення виявляється недостатнім для відображення складного розподілу між модальностями.

Багатозадачне навчання є технікою, що набула значної популярності у сфері машинного та глибокого навчання, багатоміткового навчання й багатовимірної регресії [9]. Багатозадачне навчання використовує переваги ширшого охоплення різних предметних областей шляхом одночасного вирішення кількох завдань. Такий підхід довів свою високу ефективність у багатьох сценаріях, оскільки дає змогу сформувати більш узагальнену та стійку модель. Це досягається завдяки спільному використанню знань між завданнями та зменшенню ризику перенавчання.

Відкритим питанням у багатозадачному навчанні залишається проблема балансування процесу навчання між завданнями. Поширеною практикою є призначення однакових ваг для всіх завдань або ж евристичне вагове налаштування функцій втрат кожного завдання. Перше рішення часто призводить до погіршення результатів, коли одне завдання починає домінувати у процесі навчання через надмірне значення функції втрат, що може бути зумовлене як самою функцією, так і складністю завдання. Друге рішення повністю залежить від людського судження, що обмежує його гнучкість у застосуванні до різних проблемних доменів і зазвичай потребує ресурсозатратного процесу підбору ваг.

З огляду на викладене, актуальним є **наукове завдання** щодо розроблення методів, які поєднують мультимодальне злиття з динамічним багатозадачним навчанням. Такий підхід має забезпечити ефективну інтеграцію ознак із різних модальностей та адаптивне балансування вагових коефіцієнтів завдань упродовж навчання, що дозволить уникнути домінування окремих завдань і підвищити загальну якість обробки мультимодальних даних.

Аналіз останніх досліджень і публікацій. Традиційне мультимодальне злиття зазвичай реалізується на трьох рівнях: раннє злиття, пізнє злиття та гібридне злиття.

Раннє злиття зазвичай здійснюється шляхом конкатенації необроблених або попередньо оброблених ознак кожної модальності безпосередньо після етапу вилучення ознак [10–12]. Такий підхід є простим у реалізації та потребує менш складної мережевої структури. Водночас раннє злиття стикається з труднощами у випадках, коли одна з модальностей представлена безперервним потоком даних, тоді як інша — дискретними даними. Крім того, зі зростанням кількості модальностей суттєво ускладнюється навчання міжмодальних взаємодій між гетерогенними ознаками.

Пізнє злиття передбачає використання декількох моделей для формування модально-специфічних предикативних оцінок. Надалі ці оцінки аналізуються й комбінуються для

отримання остаточного рішення [13–16]. Пізнє злиття має низку переваг порівняно з раннім. По-перше, моделі, орієнтовані на окремі модальності, дають змогу навчатися різним семантичним представленням для кожної модальності. По-друге, застосування пізнього злиття дозволяє використовувати переваги доменно-специфічних моделей і алгоритмів, наприклад, застосування моделей на основі згорткових нейронних мереж (CNN) для візуальних даних або моделей на основі рекурентних нейронних мереж (RNN) для послідовних даних. Проте пізнє злиття не враховує міжмодальні взаємодії на рівні ознак, що призводить до втрати критично важливої міжмодальної інформації.

Гібридне злиття поєднує переваги раннього та пізнього підходів, перетворюючи необроблені вхідні дані на представлення вищого рівня, що спрощує процес злиття різних модальностей і сприяє навчанню міжмодальних представлень [17; 18].

Останнім часом метод тензорного злиття привернув увагу значної кількості науковців. Тензорне злиття вирішує проблему гетерогенного розподілу даних у мультимодальному навчанні шляхом злиття кожної модальності на рівні тензора. У результаті модель здобуває можливість навчатися з урахуванням деталізації міжмодальних взаємодій. Тензорне злиття продемонструвало перспективні результати у сфері глибокого мультимодального навчання для завдань візуального питання-відповіді [19] та аналізу сентиментів [20].

Беньйоунес та інші запропонували концептуальну структуру для розв’язання задачі візуальної відповіді на запитання [19]. Вони екстрагували ознаки як із візуальних зображень, так і з текстових запитань за допомогою GRU (Gated Recurrent Unit) та ResNet [21]. Далі ознаки об’єднувалися із застосуванням методу тензорного злиття. У процесі злиття використовувався підхід тензорної декомпозиції Такера для параметризації кореляції тензорів між візуальними та текстовими представленнями. В іншій роботі Жао та інші [22] застосували мультиагентний тензорний шар і згорткове злиття для фіксації міжмодальних взаємодій.

Моделі злиття на основі графів перетворюють модальності та взаємодії між ними у графи злиття. Ознаки кожної модальності розглядаються як вершини, а взаємозв’язки між ними реалізуються у вигляді ребер. Задех та інші використали динамічний граф злиття для моделювання *n*-модальної динаміки [23]. Порівняно з тензорним злиттям графове злиття досягає вищої ефективності навчання, оскільки вимагає значно меншої кількості параметрів. Крім того, графове злиття застосовує динамічні параметри для керування активацією певних ребер, тим самим динамічно змінюючи структуру мережі. Поєднання мультимодального метричного навчання та графового злиття використовується для оцінки подібності ознак між модальностями [24]. Чен та інші [25] запропонували гетерогенну графову мережу злиття, яка орієнтується на інтеграцію мультимодальних даних із відсутніми модальностями. Вона застосовує графову мережу для проєктування відсутніх даних разом з іншими модальностями у спільний простір інтеграційного представлення.

Багатозадачне навчання забезпечує низку переваг завдяки одночасному вирішенню кількох завдань. Окрім очевидної переваги скорочення часу навчання за рахунок виконання лише одного циклу, воно також допомагає моделі формувати більш узагальнене представлення всієї предметної області, що значно знижує ризик перенавчання [9]. Багатозадачне навчання демонструє великий потенціал у сфері мультимодального навчання. Сенер та інші [26] використали багатокритеріальну оптимізацію для пошуку Парето-оптимального розв’язку шляхом мінімізації зваженої комбінації функцій втрат завдань. Ху та Сінгх [27] використали механізм “кодер-декодер”, який кодує кожну вхідну модальність та декодує їх у спільний простір інтеграційного представлення.

Мета статті – розробка методу графового злиття мультимодальних даних з використанням багатозадачного навчання, який враховуватиме міжмодальні взаємодії, адаптивно налаштовуватиме функції втрат окремих завдань та автоматично регулюватиме процес навчання моделі, як на рівні завдань, так і на рівні окремих зразків вибірки.

Виклад основного матеріалу дослідження. Враховуючи аналіз, який був проведений в [6] та сформовані завдання, запропоновано метод графового злиття мультимодальних даних з використанням багатозадачного навчання, який складається з двох основних компонентів. Перший компонент – вдосконалена графова мережа злиття (ВГМЗ), яка вирішує завдання представлення ознак кожної модальності та їх об'єднання. Другий компонент – модуль динамічного багатозадачного навчання (ДБН), який на основі спільних ознак, сформованих першим компонентом, регулює процес вирішення кожного окремого завдання.

Графова мережа злиття. Спираючись на [28], сформовано графову мережу злиття, використовуючи *n*-модальні взаємодії. Графова мережа злиття об'єднує всі модальності на унімодальному, бімодальному та тримодальному рівнях і моделює взаємодії та відношення між кожною парою комбінацій. Загальний вигляд графової мережі злиття наведено на рисунку 1.

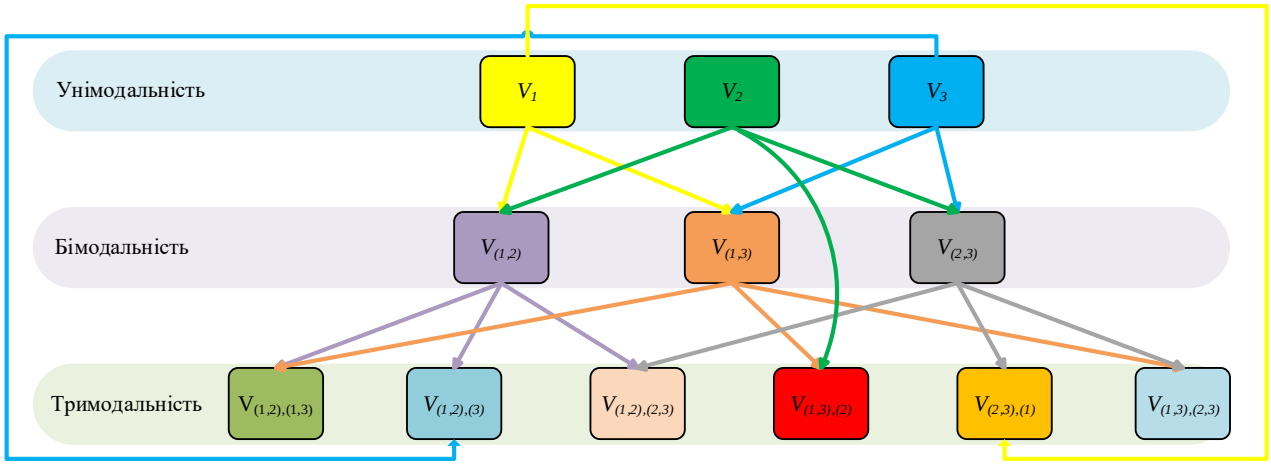

Рис. 1. Графова мережа злиття з трьома вхідними модальностями

Перший рівень графової мережі містить усі унімодальні вектори та їхні взаємодії. Унімодальний вхідний вектор ознак задано як *Vi*, де $i=[1,N]$, N – загальна кількість модальностей. Хоча графова мережа злиття може бути застосована до будь-якої кількості модальностей, надалі у статті прийнято обмеження для трьох модальностей, тобто $N=3$. Після попередньої обробки даних, застосовується паддінг для приведення векторів ознак різних модальностей до однієї розмірності. Далі вектори ознак з кожної модальності конкатенуються. Сформований вектор ознак подається до блока динамічної уваги – $\boldsymbol{B}$, який визначає значущість кожної модальності й призначає її як вагу для з'єднувальних ребер, він представлений нейронною мережею, що складається з двох згорткових шарів Conv 5×5 та Conv 1×1 і функцією активації LeakyReLU. Цей процес можна описати так:

$$w_1 \oplus w_2 \oplus \ldots \oplus w_N = \boldsymbol{B}(V_1 \oplus V_2 \oplus \ldots \oplus V_N), \quad (1)$$

де $\oplus$ – операція конкатенації; $V_1, V_2, \ldots, V_N$ – унімодальні вектори N-модальностей; w_1, w_2, w_N – відповідні ваги. $\boldsymbol{B}$ визначає динамічний коефіцієнт важливості, що має бути призначений кожному вектору у вибірково-залежний спосіб. Ці коефіцієнти важливості використовуються як основа для формування ваг ребер на вищих рівнях.

Фінальний унімодальний вектор рівня може бути отриманий як зважене середнє векторів усіх унімодальних вершин:

$$G_{unm} = \frac{1}{N}\sum_{i=1}^{N} w_i V_i. \quad (2)$$

На бімодальному рівні кожна пара унімодальних векторів об'єднується для формування вершин цього рівня. Для поєднання унімодальних векторів і побудови всіх бімодальних вершин використовується нейронна мережа – *NN*, що складається з одного одномірного згорткового шару та одного повнозв'язного шару з активацією LeakyReLU. Ця процедура описується так:

$$V_{(x,y)} = NN(V_x \oplus V_y)$$
$$\text{x=1, 2, ..., N; y=1, 2, ..., N; x}\neq\text{y,} \tag{3}$$

де $V_{(x,y)}$ – бімодальний вектор.

Щодо ребер, які з'єднують вершини між унімодальним та бімодальним рівнями, припускається, що чим ближчі дві ознаки у векторному просторі, тим більш однорідною є інформація, яку вони містять. Отже, комбінація таких ознак не надає стільки додаткової інформації, як поєднання більш відмінних ознак. Виходячи з цього припущення, обчислюється подібність між кожною парою вершин на бімодальному рівні:

$$S_{x,y} = COS(\tilde{V}_x, \tilde{V}_y), \tag{4}$$

де $S_{x,y}$ позначає показник подібності між вершинами x та y, COS – функція косинусної подібності, а $\tilde{V}_x$ та $\tilde{V}_x$ – нормовані за допомогою нормованої експоненційної функції вектори V_x та V_y. Метою нормалізації є обмеження значень обох векторів у діапазоні від 0 до 1. Відповідно до нашого припущення, чим подібніші два вектори, тим меншу вагу вони повинні мати при об'єднанні. Іншими словами, вага ребра між двома вершинами має зростати оберненопропорційно до показника подібності.

Таким чином, вага ребра, що з'єднує вершину *x* на унімодальному рівні з вершиною *xy* на бімодальному рівні, обчислюється як $\frac{w_x}{S_{x,y}+\delta}$. Аналогічно, вага ребра, що з'єднує вершину *y* та *xy*, визначається як $\frac{w_y}{S_{x,y}+\delta}$.

Показник δ є регульованим коефіцієнтом, що контролює швидкість зростання й приймає значення в інтервалі від 0 до 1. На основі емпіричних досліджень у цій роботі використано значення δ = 0.5. Відповідно, вага вершини на бімодальному рівні формалізується як:

$$q_{x,y} = \frac{w_x + w_y}{S_{x,y} + \delta},$$
$$w_{x,y} = \frac{e^{q_{x,y}}}{\sum_{j=1}^{N}\sum_{k=1,j\neq k}^{N} e^{q_{j,k}}}, \tag{5}$$

де $q_{x,y}$, є вагою вершини для $V_{(x,y)}$ на бімодальному рівні, а $w_{x,y}$ позначає нормовану за softmax форму $q_{x,y}$. Тоді остаточний об'єднаний вектор бімодального рівня можна подати так:

$$G_{bim} = \sum_{x=1}^{N}\sum_{y=1,x\neq y}^{N} w_{x,y} * V_{(x,y)}, \tag{6}$$

де G_{bim} – об'єднаний бімодальний вектор.

На тримодальному рівні всі обчислення подібні до процедури, наведеної для бімодальної частини. Формули (4)–(6) використовуються для обчислення показників подібності, ваг

вершин і об'єднаного тримодального вектора на цьому рівні. Тримодальний рівень містить два типи вершин:

комбінації бімодальних вершин;

поєднання кожної бімодальної вершини з унімодальною вершиною, яка не входила до складу цієї бімодальної вершини.

Таким чином, для набору даних із трьома вхідними модальностями загальна кількість вершин на тримодальному рівні становитиме 6.

На завершальному етапі об'єднані вектори з унімодального, бімодального та тримодального рівнів конкатенуються для формування остаточного об'єднаного вектора:

$$G_{summ} = G_{unm} \oplus G_{bim} \oplus G_{trm}. \quad (7)$$

Динамічне багатозадачне навчання. У багатозадачному навчанні підсумкова функція втрат обчислюється як лінійна комбінація втрат усіх завдань, що використовується для оптимізації параметрів моделі. Типові підходи багатозадачного навчання або призначають усім завданням однакові ваги, або ж визначають вагу для кожного завдання на основі емпіричних досліджень. В роботі запропоновано модуль динамічного багатозадачного навчання, здатний обчислювати та корегувати ваги функцій втрат як на рівні вибірки, так і на рівні завдань. Крім того, модуль здійснює повторне балансування початкових значень функцій втрат на кожному циклі навчання, аби уникнути потрапляння вагових коефіцієнтів у локальні мінімуми чи максимуми.

Модуль динамічного багатозадачного навчання на рівні вибірки має на меті надавати вищий пріоритет навчанню тих вхідних прикладів, які були класифіковані неправильно. Використовуючи кросс-ентропію, як функцію втрат, цей процес можна описати так:

$$H(p_d) = -\log(p_d), \quad (8)$$

$$p_d = \begin{cases} p, & \text{якщо } y = 1 \\ 1 - p, & \text{інакше} \end{cases}, \quad (9)$$

де $y \in \{0,1\}$ – істинна мітка, $a\ 0 \le p \le 1$ – ймовірність того, що зразок має мітку 1. Функція вагового коефіцієнта втрат на рівні вибірки L_s визначається як:

$$L_s(x) = -(1 - p_d)^{\beta} \log(p_d), \quad (10)$$

де x — це вхідні дані;

β — параметр фокусування на рівні вибірки, який контролює величину зменшення ваги для простих (істинно-негативних) зразків.

На основі емпіричних досліджень у цій роботі використано значення $\beta = 2$. Коли p_d є малим і зразок класифіковано неправильно, значення $(1 - p_d)^{\beta}$ наближається до 1, що має дуже обмежений вплив на функцію втрат. З іншого боку, зі збільшенням p_d величина $(1 - p_d)^{\beta}$ поступово прямує до 0, що зменшує втрати, породжені правильно класифікованими зразками. Це змушує модель приділяти більше ресурсів навчанню складних (хибно-негативних) зразків.

Для порівняння, модуль динамічного багатозадачного навчання на рівні завдань автоматично коригує ваги втрат, згенерованих кожною мережею, що відповідає за певне завдання. Функція втрат кожного завдання контролюється та використовується як метрика для регулювання зваженого градієнта в кожному шарі. В роботі застосовується спеціалізована функція втрат для мінімізації різниці між: градієнтами ваг серед усіх завдань та зваженим за

швидкістю навчання середнім градієнтом. L1-норма динамічної функції балансування втрат на рівні завдань, на ітерації навчання *t* визначається так:

$$L_D(t) = \sum_f \frac{M}{m_f} \left| G_W^{(f)}(t) - \bar{G}_W(t) \times \left[r_f(t) \right]^{\alpha} \right|_1, \quad (11)$$

де M представляє загальну кількість вибірок,

m_f – кількість позитивних зразків у завданні *f*;

$\frac{M}{m_f}$ – обернене співвідношення розподілу вибірки для завдання *f*;

W – ваговий параметр останнього шару мережі, специфічної для завдання;

$G_W^{(f)}(t)$ – *L2*-норма градієнта зваженої функції втрат завдання *f* на ітерації *t*;

$\bar{G}_W(t)$ – середній градієнт усіх завдань на ітерації *t*;

$r_f(t)$ – обернена швидкість навчання завдання *f*;

α – коефіцієнт, що контролює величину оберненої швидкості навчання.

У деяких випадках оновлення ваг функції втрат через обчислення динамічної функції балансування втрат може бути недостатнім, якщо складність завдань є надто нерівномірною. Це може уповільнити процес оновлення функцій втрат і призвести до потрапляння ваг деяких завдань у локальні мінімуми чи максимуми. Щоб розв'язати цю проблему, виконується повторне балансування ваг функцій втрат після завершення повного циклу навчання. Це забезпечує більш агресивний процес оновлення втрат, що допомагає моделі швидше досягти оптимальних ваг функцій втрат і уникнути їх потрапляння в локальні мінімуми чи максимуми.

На практиці середні значення функцій втрат для кожного завдання обчислюються протягом усього циклу навчання. Потім ваговий коефіцієнт генерується шляхом ділення найбільшого значення серед усіх середніх втрат на середню втрату кожного завдання. Нарешті, цей ваговий коефіцієнт застосовується до всіх завдань для повторного балансування втрат на початку циклу навчання. Відстежується втрата на валідаційній вибірці, і процес навчання зупиняється, якщо ця втрата перестає зменшуватися.

Оцінка ефективності запропонованого методу

CrisisMMD [29] – це мультимодальний мультимедійний набір даних із Twitter, який містить понад 16000 твітів і 18000 зображень, пов'язаних із сімома великими природними катастрофами. Кожен зразок анотовано за трьома групами концептів: інформаційний рівень даних, гуманітарна категорія та рівень руйнувань. Інформаційний рівень даних відображає обсяг корисної інформації, гуманітарна категорія охоплює тип гуманітарної кризи та заходи з ліквідації наслідків, що відбувалися на місці події, а рівень руйнувань – характеризує масштаб пошкоджень інфраструктури та комунальних об'єктів. Для цього набору даних подаються результати за метриками F1, HL та MAP. Для метрик F1 та MAP більші значення означають кращу якість моделі, тоді як для HL – нижчі значення є кращими.

Виділення візуальних ознак. Для обробки візуальних даних використовується методика InceptionV3 [30], попередньо навчена на наборі ImageNet [31], як екстрактор ознак.

Виділення текстових ознак. Для текстових даних використовується методика ELMo [32] для формування векторних уявлень слів. Порівняно з традиційними методами текстової векторизації, такими як Word2vec [33] та GloVe [34], ELMo здатний враховувати морфологічну інформацію та ефективніше працює з позасловниковими словами.

Розподіл даних для експерименту: 60 % даних використовується для навчання, 20 % – для валідації та 20 % – для тестування. Набір для валідації застосовується для налаштування всіх гіперпараметрів, а коефіцієнт α у динамічній функції втрат встановлюється рівним 1 на

основі емпіричних досліджень. Для оптимізації процесу навчання використовується алгоритм Adam [35], а початкова швидкість навчання встановлена на рівні 0,01.

Модуль динамічного багатозадачного навчання застосовується до трьох груп концептів у наборі даних CrisisMMD, де кожен концепт моделюється як окреме завдання. Це перетворює початкову задачу багатоміткової класифікації на задачу багатозадачного навчання.

Результати та обговорення. Для демонстрації ефективності запропонованої структури було обрано кілька базових методів, включно з найсучаснішими підходами.

Базові методи мультимодального злиття включають:

метод конкатенаційного злиття (КЗ), який виконує просту конкатенацію ознак кожної модальності безпосередньо після етапу початкового виділення ознак;

тензорний метод злиття (ТМЗ) [21];

графова мережа злиття (ГМЗ) [27].

Базові методи для багатозадачного навчання включають:

лінійна сума функцій втрат із рівними вагами (ЛСРВ);

багатооб'єктна оптимізація (БОО) [25];

багатомасштабні мережі взаємодії завдань (БМВЗ) [36].

Для цілей порівняння у кожному базовому методі нижчі рівні моделі замінено на вищезгадані попередньо навчені моделі.

Мультимодальне злиття. У таблицях 1, 2 та 3 наведено результати ефективності запропонованого методу, а також базових методів для концептів “інформаційний рівень даних”, “гуманітарна категорія” та “рівень руйнувань” у наборі даних CrisisMMD.

Таблиця 1

Оцінювання ефективності за інформаційним концептом даних у наборі CrisisMMD

Методи	F1	HL	MAP
КЗ + ЛСРВ	0,633	0,227	0,597
ТМЗ + ЛСРВ	0,784	0,151	0,738
ГМЗ + ЛСРВ	0,823	0,114	0,772
ВГМЗ + ЛСРВ	0,849	0,107	0,794
КЗ + БОО	0,675	0,202	0,648
КЗ + БМВЗ	0,683	0,214	0,635
КЗ + ДБН	0,746	0,164	0,719
ВГМЗ + ДБН	**0.872**	**0,041**	**0,835**

Таблиця 2

Оцінювання ефективності за концептом гуманітарної категорії у наборі CrisisMMD

Методи	F1	HL	MAP
КЗ + ЛСРВ	0,537	0,293	0,506
ТМЗ + ЛСРВ	0,691	0,207	0,659
ГМЗ + ЛСРВ	0,687	0,209	0,652
ВГМЗ + ЛСРВ	0,722	0,181	0,695
КЗ + БОО	0,613	0,246	0,581
КЗ + БМВЗ	0,624	0,237	0,598
КЗ + ДБН	0,696	0,194	0,670
ВГМЗ + ДБН	**0,772**	**0,153**	**0,759**

Таблиця 3

Оцінювання ефективності за концептом рівня руйнувань у наборі CrisisMMD

Методи	F1	HL	MAP
КЗ + ЛСРВ	0,644	0,229	0,607
ТМЗ + ЛСРВ	0,791	0,148	0,745
ГМЗ + ЛСРВ	0,829	0,117	0,793
ВГМЗ + ЛСРВ	0,862	0,080	0,839

Методи	F1	HL	MAP
КЗ + БОО	0,693	0,181	0,664
КЗ + БМВЗ	0,688	0,186	0,650
КЗ + ДБН	0,756	0,149	0,715
ВГМЗ + ДБН	**0,923**	**0,029**	**0,897**

Як видно комбінація КЗ + ЛСРВ показує найнижчі результати за всіма метриками. Це не дивно, оскільки проста конкатенація ознак на ранньому етапі часто не здатна відобразити гетерогенний розподіл різних модальностей. Крім того, лінійна сума функцій втрат із рівними вагами в межах багатозадачного навчання має дуже обмежену ефективність або навіть негативний вплив, коли кілька завдань домінують у процесі навчання.

Тензорний метод злиття та графовий метод злиття – обидва демонструють покращення ефективності порівняно з підходом КЗ + ЛСРВ. При цьому ГМЗ має помітну перевагу над ТМЗ, особливо у випадках з більшою кількістю вхідних модальностей. Це частково пояснюється тим, що традиційні методи тензорного злиття на кшталт ТМЗ моделюють спільне представлення лише після виконання операції злиття, тоді як ГМЗ долає цей недолік, навчаючись міжмодальним взаємодіям уже на ранніх етапах.

Запропонований нами вдосконалений графовий метод злиття перевершує всі базові методи й перевищує результат другого найкращого підходу на **4,2 % за метрикою F1** та на **8,5 % за MAP**. Ми вважаємо, що частково це зумовлено застосуванням динамічного показника **B**, який навчається визначати відносну важливість кожної модальності та інтегрує її вже на початковому етапі графової мережі злиття.

Багатозадачне навчання. У таблицях 1–3 також наведено результати роботи методів багатозадачного навчання на наборі даних CrisisMMD. Методи БОО та БМВЗ демонструють кращу ефективність порівняно з підходом багатозадачного навчання із лінійною сумою функцій втрат з рівними вагами. Проте загальне покращення не є суттєвим. Ймовірним поясненням цього є ситуація дисбалансності класів, коли обидва методи зазнають значного зниження ефективності.

Запропонований нами метод вирішує проблему дисбалансу класів завдяки введенню у функцію втрат параметра оберненого співвідношення вибірки для завдань. Це дозволяє моделі посилювати штрафування класів-більшостей, виділяючи більше ресурсів на навчання класів-меншостей. Крім того, ми стверджуємо, що повторне балансування ваг функцій втрат на початку циклу навчання допомагає моделі й надалі зменшувати загальну навчальну втрату, тоді як у двох інших методів цей механізм відсутній.

Для набору даних CrisisMMD запропонований метод перевищує другий найкращий метод на **7,2 % за F1** та на **7,3 % за MAP**.

Загалом, запропонований нами метод з графовою мережею злиття та динамічним багатозадачним навчанням демонструє найкращі результати серед усіх базових методів при роботі з мультимодальним набором даних **CrisisMMD**. Крім того, модульність компонентів методу забезпечує високу гнучкість та простоту їхнього застосування до інших типів даних і структур моделей.

Висновки. У цій роботі запропоновано новий метод графового злиття мультимодальних даних з використанням багатозадачного навчання. Запропонований метод здійснює ієрархічне поєднання кожної модальності з утворенням графової структури, де вершини відповідають об'єднаним модальностям, а ребра містять міжмодальні взаємодії. Відносна важливість між модальностями на одному рівні визначається у вибірково-орієнтованій манері та використовується для формування спільного інтеграційного представлення, яке надалі передається на наступний рівень.

Крім того, ми пропонуємо новий підхід динамічного багатозадачного навчання, що трансформує задачу багатоміткової класифікації у сукупність окремих двійкових задач

класифікації. Завдяки моніторингу складності навчання для кожного завдання компонент динамічного багатозадачного навчання автоматично коригує вагові коефіцієнти функції втрат завдань так, щоб досягався оптимальний баланс. Цей компонент також призначає початковий набір вагових коефіцієнтів для функцій втрат на початку циклу навчання та постійно оновлює їх упродовж процесу аби уникнути потрапляння у локальні мінімуми чи максимуми.

Експериментальні результати на мультимодальному наборі даних продемонстрували суттєву перевагу нашого методу над базовими методами. Більше того, запропонований метод завдяки своїй модульній архітектурі може бути легко застосований до інших типів даних і архітектур мереж.

Таким чином отриманий метод є підґрунтям для подальшої розробки методики обробки мультимодальних даних в інформаційних системах військового призначення, яка дасть можливість підвищити ефективність обробки великих масивів даних та якість прийняття управлінських рішень.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Dashenkov D., Smelyakov K. Extending the ImageNET dataset for multimodal text and image learning // INNOVATIVE TECHNOLOGIES AND SCIENTIFIC SOLUTIONS FOR INDUSTRIES. 2025. No. 1 (31)/. P. 20–31. URL: https://doi.org/10.30837/2522-9818.2025.1.020.

2. Олег Басистюк, Наталія Мельникова, Ірина Думин, Андрій Думин. Ансамблевий підхід у мультимодальній обробці даних на основі Google API // Вісник Хмельницького національного університету. Серія: Технічні науки. 2024. № 4 (339). С. 235–238.

3. Yuan Tao, Wanzeng Liu, Jun Chen, Jingxiang Gao, Ran Li, Xinpeng Wang, Ye Zhang, Jiaxin Ren, Shunxi Yin, Xiuli Zhu, Tingting Zhao, Xi Zhai, Yunlu Peng. A graph-based multimodal data fusion framework for identifying urban functional zone // International Journal of Applied Earth Observation and Geoinformation. 2025. Vol. 136. URL: https://doi.org/10.1016/j.jag.2024.104353.

4. Dong J., Hao M., Ding F., Chen S., Wu J., Zhuo J., Jiang D. A Novel Multimodal Data Fusion Framework: Enhancing Prediction and Understanding of Inter-State Cyberattacks // Big Data and Cognitive Computing. 2025. No. 9 (3)/. P. 63. URL: https://doi.org/10.3390/bdcc9030063.

5. T. Baltruˇ saitis, C. Ahuja, L.-P. Morency. Multimodal machine learning: A survey and taxonomy // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2018. Vol. 41, no. 2. Pp. 423–443.

6. Меркотан Д. Ю., Троцько О. О. Методики обробки мультимодальних даних в інформаційних системах з використанням моделей штучного інтелекту // Збірник наукових праць Харківського національного університету Повітряних Сил. 2025. № 2 (84). С. 117–125. URL: https://doi.org/10.30748/zhups.2025.84.14.

7. P. Hu, Y.-A. Huang, K. C. Chan, Z.-H. You. Learning multi modal networks from heterogeneous data for prediction of lncrna–mirna interactions // IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2019. Vol. 17, no. 5. Pp. 1516–1524.

8. C. G. Snoek, M. Worring, A. W. Smeulders. Early versus late fusion in semantic video analysis // in Proceedings of the 13th Annual ACM International Conference on Multimedia. 2005. Pp. 399–402.

9. Y. Zhang, Q. Yang. A survey on multi-task learning // IEEE Transactions on Knowledge and Data Engineering, 2021.

10. Liu Z., Yin Z., Mi Z., Guo B., Zheng Z. A Comparative Analysis of Three Data Fusion Methods and Construction of the Fusion Method Selection Paradigm // Mathematics. 2025. No. 13 (8). P. 1218. URL: https://doi.org/10.3390/math13081218.

11. Songtao L., Hao T. Multimodal Alignment and Fusion: A Survey, arXiv:2411.17040v1 [cs.CV]. 26 Nov 2024.

12. K. Gadzicki, R. Khamsehashari, C. Zetzsche. Early vs Late Fusion in Multimodal Convolutional Neural Networks // 2020 IEEE 23rd International Conference on Information Fusion (FUSION), Rustenburg, South Africa, 2020. Pp. 1–6. DOI: 10.23919/FUSION45008.2020.9190246.

13. Y. Zhang, O. Morel, M. Blanchon, R. Seulin, M. Rastgoo, D. Sidib´e. Exploration of deep learning-based multimodal fusion for semantic road scene segmentation in VISIGRAPP (5: VISAPP). 2019. Pp. 336–343.

14. Qihang Yang, Yang Zhao, Hong Cheng. MMLF: Multi-modal Multi-class Late Fusion for Object Detection with Uncertainty Estimation // arXiv:2410.08739v1. URL: https://doi.org/10.48550/arXiv.2410.08739.

15. Clemens Witt, Thiemo Leonhardt, Nadine Bergner, Mareen Grillenberger. Multimodal Late Fusion Model for Problem-Solving Strategy Classification in a Machine Learning Game // arXiv:2507.22426v1 [cs.LG]. 30 Jul 2025.

16. Y. Cheng, R. Cai, Z. Li, X. Zhao, K. Huang. Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation // in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017., Pp. 3029–3037.

17. J. Jiang, L. Zheng, F. Luo, Z. Zhang. Rednet: Residual encoder decoder network for indoor RGB-D semantic segmentation // CoRR. 2018. Vol. abs/1806.01054.

18. A. Valada, R. Mohan, W. Burgard. Self-supervised model adap tation for multimodal semantic segmentation // International Journal of Computer Vision. 2019. Pp. 1–47.

19. H. Ben-Younes, R. Cadene, M. Cord, N. Thome. Mutan: Multimodal tucker fusion for visual question answering // in Proceedings of the IEEE International Conference on Computer Vision. 2017. Pp. 2612–2620.

20. A. Zadeh, M. Chen, S. Poria, E. Cambria, L. Morency. Tensor fusion network for multimodal sentiment analysis // in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics. 2017. Pp. 1103–1114.

21. Desai A., Mahto R. Multi-Class Classification of Breast Cancer Subtypes Using ResNet Architectures on Histopathological Images // Journal of Imaging. 2025. No. 11 (8). P. 284. URL: https://doi.org/10.3390/jimaging11080284.

22. T. Zhao, Y. Xu, M. Monfort, W. Choi, C. Baker, Y. Zhao, Y. Wang, Y. N. Wu. Multi-agent tensor fusion for contextual trajectory prediction // in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. Pp. 12126–12134.

23. A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, L.-P. Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph // in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018. Pp. 2236–2246.

24. M. Angelou, V. Solachidis, N. Vretos, P. Daras. Graph-based multimodal fusion with metric learning for multimodal classification // Pattern Recognition. 2019. Vol. 95. Pp. 296–307.

25. J. Chen, A. Zhang. Hgmf: Heterogeneous graph-based fusion for multimodal data with incompleteness // in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020. Pp. 1295–1305.

26. O. Sener, V. Koltun. Multi-task learning as multi-objective optimization // in Annual Conference on Neural Information Processing Systems. 2018. Pp. 525–536.

27. R. Hu, A. Singh. Unit: Multimodal multitask learning with a unified transformer // arXiv preprint arXiv:2102.10772. 2021.

28. S. Mai, H. Hu, S. Xing. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion // in Proceedings of the AAAI Conference on Artificial Intelligence. 2020. Vol. 34, no. 01. Pp. 164–172.

29. F. Alam, F. Ofli, M. Imran. Crisismmd: Multimodal twitter datasets from natural disasters // in Proceedings of the International AAAI Conference on Web and Social Media. 2018. Vol. 12, no. 1.

30. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna. Rethinking the inception architecture for computer vision // in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. Pp. 2818–2826.

31. J. Deng, W. Dong, R. Socher, L. Li, K. Li, F. Li. Imagenet: A large-scale hierarchical image database // in IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society. 2009. Pp. 248–255.

32. M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer. Deep contextualized word representations // in Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT). Association for Computational Linguistics. 2018. Pp. 2227–2237.

33. T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient estimation of word representations in vector space // in 1st International Conference on Learning Representations (ICLR), Y. Bengio and Y. LeCun, Eds., 2013.

34. J. Pennington, R. Socher, C. D. Manning. Glove: Global vectors for word representation // in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). ACL, 2014. Pp. 1532–1543.

36. D. P. Kingma, J. Ba. Adam: A method for stochastic optimization // in 3rd International Conference on Learning Representations (ICLR), 2015.

37. S. Vandenhende, S. Georgoulis, L. Van Gool. Mti-net: Multi-Scale task interaction networks for multi-task learning // in European Conference on Computer Vision. Springer. 2020. Pp. 527–543.