УДК 004.8:004.93:004.91

ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ АНАЛІЗУ ТА КЛАСИФІКАЦІЇ МУЛЬТИМЕДІЙНОЇ ІНФОРМАЦІЇ

[1]Олександр СУПРУН, [2]Іван ЖИЛЯК
[1] канд. техн. наук, доцент, кафедра МСТ, ХНУРЕ
[2]студент, кафедра ЕОМ, ХНУРЕ

Abstract. *The paper examines intelligent systems for analysis and classification of multimedia information, which combine methods of computer vision, natural language processing, and audio analysis. Approaches to unimodal and multimodal data processing are systematized, fusion strategies (early, late, and intermediate) are compared, and a methodology for evaluating classification quality based on the F1-measure is presented. The advantages of cross-modal transformer architectures, their application areas in publishing and media industries, as well as key challenges — data heterogeneity, missing modalities, and model interpretability — are analyzed.*

Keywords: multimedia information, intelligent analysis, classification, deep learning, multimodal fusion, transformer, F1-measure.

Стрімке зростання обсягів цифрового контенту текстів, зображень, аудіо- та відеофайлів — зумовлює потребу в інтелектуальних системах, здатних автоматично аналізувати та класифікувати мультимедійну інформацію. За оцінками дослідників, понад 80% даних, що генеруються у світі, мають мультимедійну природу, і більша їх частина залишається неструктурованою. Традиційні інформаційні системи, орієнтовані на обробку окремих типів даних, не в змозі виявляти комплексні закономірності, що проявляються лише при спільному розгляді різних модальностей, тому на перший план виходять мультимодальні підходи до аналізу [1]. Мультимедійна інформація характеризується гетерогенністю модальностей, які суттєво відрізняються за природою, структурою та обсягом. Інтелектуальна система аналізу повинна забезпечувати попередню обробку кожної модальності (нормалізацію, видалення шуму, сегментацію), витягування інформативних ознак та їхнє подальше об’єднання для прийняття класифікаційного рішення [2].

Для аналізу окремих модальностей сформувалися сталі архітектурні рішення. У галузі обробки зображень домінують згорткові нейронні мережі (CNN) та Vision Transformer, які послідовно витягають ознаки від локальних текстур до семантичних концепцій; огляди останніх років засвідчують, що глибинне навчання забезпечує точність класифікації зображень, близьку до людської, на репрезентативних вибірках [3].

Процедура класифікації полягає у віднесенні об’єкта до одного з наперед визначених класів на підставі витягнутих ознак. Класичні методи машинного навчання опорні вектори (SVM), випадкові ліси, k-найближчих сусідів залишаються ефективними на задачах із малим обсягом навчальних даних і високорозмірними ознаками, проте їхня якість критично залежить від ручного конструювання ознак. Натомість глибинні моделі виконують витягування ознак і класифікацію як єдиний диференційовий процес, що забезпечує значно вищу узагальнювальну здатність за наявності великих наборів даних [4].

Ключовим питанням інтелектуального аналізу мультимедійної інформації є стратегія злиття модальностей. У ранній фузії ознаки всіх модальностей

конкатенуються на рівні вхідних векторів, що дозволяє моделі вчити спільні представлення, але чутливе до пропусків даних. Пізня фузія передбачає незалежне навчання моделей для кожної модальності з подальшим об'єднанням їхніх прогнозів — такий підхіст стійкіший до втрати модальностей. Проміжні та крос-модальні схеми, зокрема архітектури на основі трансформерів з механізмами уваги, забезпечують взаємне збагачення представлень: текстова модальність «підказує», на які фрагменти зображення звернути увагу, і навпаки.

Порівняльний аналіз підходів (рис. 1) показує, що унімодальні системи демонструють точність на рівні 83–91%, тоді як мультимодельні архітектури з крос-модальною увагою досягають 96% на задачах аналізу складного контенту. Найбільшого приросту якості (на 5–10 процентних пунктів) досягають системи, що об'єднують текстову та візуальну модальності, саме ця комбінація є базовою для задач автоматичної модерації контенту, виявлення дезінформації та семантичного індексування мультимедійних архівів.

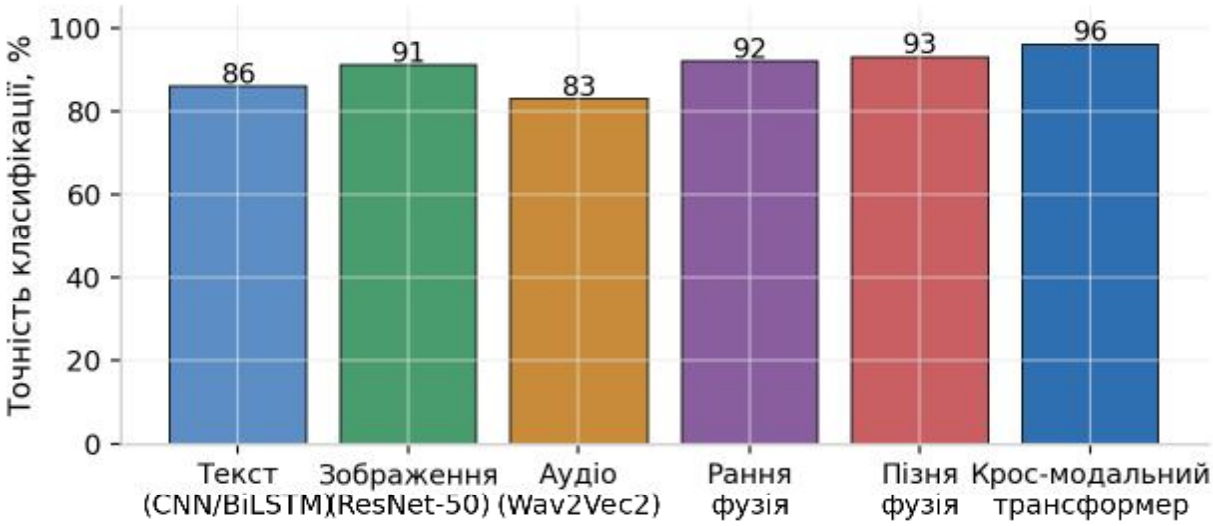

Рис. 1. Порівняння точності класифікації унімодальних та мультимодальних підходів

Оцінювання якості класифікації в умовах неврівноважених класів доцільно здійснювати за узагальненою F-мірою, що гармонійно поєднує точність (precision) і повноту (recall):

$$F_1 = 2 \cdot P \cdot R / (P + R), \quad (1)$$

де P — точність, частка об'єктів, правильно віднесених до класу серед усіх об'єктів, віднесених моделлю до цього класу; R — повнота, частка об'єктів класу, виявлених системою; F_1 — гармонійне середнє, яке дорівнює 1 лише за ідеальної точності та повноти і суттєво штрафує моделі, що досягають високого значення лише одного з показників. У практичних системах аналізу мультимедійного контенту поріг прийнятності зазвичай встановлюється на рівні $F_1 \geq 0{,}9$.

Незважаючи на вражаючі результати, інтелектуальні системи аналізу мультимедійної інформації натрапляють на низку викликів. Гетерогенність модальностей ускладнює узгодження їхніх представлень у спільному просторі; реальні дані часто містять пропущені або пошкоджені модальності; великі моделі

вимагають значних обчислювальних ресурсів, що обмежує їхнє впровадження на кінцевих пристроях. Окремої уваги потребує інтерпретовність: методи пояснювального штучного інтелекту (XAI) для мультимодальних моделей візуалізація уваги, аналіз внеску модальностей лише починають розвиватися, хоча саме вони забезпечують довіру до рішень системи у критичних застосуваннях.

Практичне значення інтелектуальних систем класифікації мультимедійної інформації виходить далеко за межі академічних досліджень. У видавничій справі та поліграфії вони забезпечують автоматичну каталогізацію ілюстративного матеріалу, верифікацію авторських прав і контроль відповідності вмісту цільовій аудиторії. У медіаіндустрії модерацію користувацького контенту та виявлення маніпулятивних публікацій, що поєднують текст і зображення. У медицині класифікацію діагностичних даних за знімками, клінічними описами та лабораторними показниками.

Універсальність підходів на основі мультимодальних трансформерів робить їх ключовою технологією наступного етапу цифрової трансформації галузей, пов’язаних з обробкою контенту. Вибір стратегії фузії та методу оцінювання якості має здійснюватися з урахуванням специфіки даних і вимог до надійності системи. Подальші дослідження доцільно спрямувати на розробку універсальних архітектур, стійких до пропусків модальностей, підвищення інтерпретовності рішень та зниження обчислювальної вартості моделей для їхнього впровадження в ресурсообмежених середовищах.

Список літератури

1. Barua, A., Ahmed, M. U., & Begum, S. (2023). A systematic literature review on multimodal machine learning: *Applications, challenges, gaps and future directions. IEEE Access, 11,* 14804–14831. https://doi.org/10.1109/ACCESS.2023.3243854

2. Liu, T., Hu, Y., Gao, J., Sun, Y., & Yin, B. (2024). Hierarchical multi-modal transformer for cross-modal long document classification [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2407.10105

3. Stahlschmidt, S. R., Ulfenborg, B., & Synnergren, J. (2022). *Multimodal classification: Current landscape, taxonomy and future directions. ACM* Computing Surveys, 54(10s), 1–38. https://doi.org/10.1145/3543848

4. Wang, B., Feng, Y., Xiong, X.-C., Qiang, Y.-H., & Qiang, B.-H. (2023). Multi-modal transformer using two-level visual features for fake news detection. *Applied Intelligence, 53(9),* 10429–10443. https://doi.org/10.1007/s10489-023-04548-6