

**РОЗРОБКА МЕТОДУ ПОЯСНЮВАНOSTІ
МУЛЬТИМОДАЛЬНИХ НЕЙРОННИХ МОДЕЛЕЙ
У ЗАДАЧАХ КЛАСИФІКАЦІЇ НЕОДНОРІДНИХ ДАНИХ**

Рудой В.В.

email: valerii.rudoi@nure.ua

Науковий керівник – д.т.н., професор Мінухін С.В.

Харківський національний університет радіоелектроніки, каф. КМІТ
м. Харків, Україна

This research focuses on enhancing the explainability of multimodal neural networks for classification of heterogeneous data. We propose a method to quantify the contribution of each modality and individual features to the final prediction. The approach combines modality exclusion analysis with gradient-based feature importance assessment. It generates structured explanations at both modality and feature levels without modifying the model architecture. The method is applicable to diverse multimodal neural architectures and integration strategies. Experimental evaluation demonstrates its effectiveness in improving transparency and interpretability of complex classification systems.

Зростання обсягів різномірних даних у сучасних інформаційних системах зумовлює необхідність застосування методів їх комплексного аналізу. У багатьох прикладних задачах інформація представлена різними типами даних, зокрема зображеннями, текстами, числовими характеристиками або сигналами. Такі дані називають мультимодальними, оскільки вони формуються з кількох джерел і мають різну структуру представлення. Використання мультимодальних підходів дозволяє підвищити якість класифікації, оскільки різні модальності містять взаємодоповнюючу інформацію про об'єкт дослідження [1].

Для обробки мультимодальних даних широко застосовуються нейронні мережі глибокого навчання. У загальному випадку мультимодальна модель виконує перетворення кожної модальності у вектор ознак, після чого формується їх спільне представлення. Нехай набір модальностей представлений множиною:

$$M = \{m_1, m_2, \dots, m_k\}, \quad (1)$$

де m_i – окрема модальність даних. Для кожної модальності використовується окрема функція відображення:

$$h_i = f_i(m_i), \quad (2)$$

де $f_i(\cdot)$ – нейронна підмережа, що виконує виділення ознак, а h_i – вектор латентного представлення відповідної модальності [2]. Отримані представлення об'єднуються у спільний вектор ознак:

$$z = \phi(h_1, h_2, \dots, h_k), \quad (3)$$

де $\phi_i(\cdot)$ – функція інтеграції модальностей. У простішому випадку вона може бути реалізована як конкатенація векторів:

$$z = [h_1 \parallel h_2 \parallel \dots \parallel h_k]. \quad (4)$$

Після формування спільного представлення виконується класифікація об'єкта за допомогою функції:

$$\hat{y} = g(z), \quad (5)$$

де $g(\cdot)$ – класифікаційна нейронна мережа, $\hat{y}$ – прогнозований клас об'єкта.

Незважаючи на високу точність мультимодальних моделей, їх використання супроводжується проблемою низької інтерпретованості результатів. Більшість сучасних нейронних моделей функціонують як «чорні скриньки», що ускладнює аналіз причин прийняття рішення. Це створює суттєві обмеження для використання таких систем у задачах, де необхідне пояснення результатів класифікації [3].

Для розв'язання цієї проблеми у роботі запропоновано метод пояснюваності, який дозволяє оцінювати внесок кожної модальності у формування результату класифікації. Нехай вихідна функція класифікації має вигляд:

$$F(M) = \{m_1, m_2, \dots, m_k\}. \quad (6)$$

Тоді внесок окремої модальності m_i можна оцінити через зміну результату класифікації при її виключенні:

$$I_i = |F(m_1, \dots, m_i, \dots, m_k) - F(m_1, \dots, 0, \dots, m_k)|, \quad (7)$$

де I_i – показник важливості модальності m_i . Чим більше значення заданого показника, тим більший вплив відповідної модальності на кінцевий результат.

Для забезпечення порівнянності значень вводиться нормалізований показник важливості:

$$\tilde{I}_i = \frac{I_i}{\sum_{j=1}^k I_j}, \quad (8)$$

що дозволяє визначити відносний внесок кожної модальності у процес класифікації.

Крім оцінювання внеску модальностей, метод передбачає аналіз важливості окремих ознак у межах кожної модальності. Для цього використовується градієнтний підхід, що дозволяє визначити чутливість вихідної функції до змін вхідних параметрів:

$$S_j = \left| \frac{\partial F}{\partial x_j} \right|, \quad (9)$$

де x_j – окрема ознака вхідних даних, S_j – її значущість для класифікації [4].

Таким чином формується пояснення результату класифікації, яке включає два рівні інтерпретації:

- внесок кожної модальності у формування рішення;
- важливість окремих ознак у межах кожної модальності.

Запропонований метод дозволяє отримувати структуроване пояснення роботи мультимодальної моделі без зміни її архітектури. Це дає можливість застосовувати його до різних типів нейронних мереж та різних стратегій інтеграції модальностей.

Перевагою підходу є універсальність і можливість використання у задачах аналізу складних неоднорідних даних. Отримані пояснення дозволяють визначати причини помилкових класифікацій, оцінювати якість використання різних модальностей та підвищувати надійність інтелектуальних систем.

Запропонований метод може бути використаний у системах комп'ютерного зору, аналізу мультимедійних даних, медичної діагностики та інших задачах, де необхідно поєднувати дані різної природи та забезпечувати інтерпретованість результатів роботи нейронних моделей [5].

Список використаних джерел:

1. Minukhin S., Rudoi V. Research on neural network architectures for multimodal data processing. Concepts for the Development of Society's Scientific Potential: proceedings of the 8th International Scientific and Practical Conference, Prague, Czech Republic, Prague. 2025. P. 214–222. DOI: <https://doi.org/10.51582/interconf.19-20.10.2025.024>
2. Rudoi V. Optimization of multimodal neural networks using attention mechanisms for the classification of diabetes stages. Débats scientifiques et orientations prospectives du développement scientifique : proceedings of the IX International Scientific and Practical Conference, Paris, République française, October 2025. Paris. 2025. P. 123–129. DOI: <https://doi.org/10.36074/logos-31.10.2025.022>
3. Wang W., Bao H. Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks. arXiv. 2024. DOI: <https://doi.org/10.48550/arXiv.2208.10442>
4. Huang S., et al. Language Is Not All You Need: Aligning Perception with Language Models. arXiv preprint. 2023. 25 p. DOI: <https://doi.org/10.48550/arXiv.2302.14045>
5. Nykonnyk M., Basystiuk O., Shakhovska N., Melnykova N. Multimodal Data Fusion for Depression Detection Approach. Computation. 2025. Vol. 13, Issue 1. P. 9. DOI: <https://doi.org/10.3390/computation13010009>