Жеребецький Олег Вячеславович, аспірант каф. СШІ

Національний університет “Львівська політехніка”, м. Львів

Шамуратов Олексій Юрійович, д.ф., асистент каф. СШІ

Національний університет “Львівська політехніка”, м. Львів

**МУЛЬТИМОДАЛЬНИЙ АНАЛІЗ ДАНИХ ДЛЯ РОЗПІЗНАВАННЯ ЕМОЦІЙ**

**Вступ**

Швидкий розвиток штучного інтелекту та машинного навчання дозволяє аналізувати людські емоції за допомогою мультимодальних даних [1]. Розпізнавання емоцій через інтонацію мови, вираз обличчя та текстові повідомлення дає змогу створювати інтелектуальні системи для оцінки психологічного стану користувачів. Враховуючи зростаючий рівень стресу через зовнішні фактори, такі як війна, економічна нестабільність та соціальний тиск, потреба в автоматизованих інструментах для моніторингу емоційного стану значно зросла. Багато людей неохоче звертаються до психологів через питання конфіденційності або дискомфорт під час спілкування зі спеціалістами. Система мультимодального розпізнавання емоцій може слугувати цифровим асистентом, що надає психологічні оцінки на основі комбінації вхідних даних. Під мультимодальними даними потрібно розуміти набір таких вхідних даних, який містить відео, аудіо та текст.

Метою цього дослідження є розробка системи, що інтегрує мультимодальні дані для розпізнавання емоцій, покращуючи точність класифікації порівняно з одномодальними підходами. Дослідження охоплює різні архітектури нейронних мереж, включаючи CNN[2], LSTM[3] та трансформери, для визначення оптимальної конфігурації обробки мультимодальних даних. Також оцінюється вплив різних наборів даних та методів попередньої обробки на точність і надійність моделей розпізнавання емоцій.

**1. Мультимодальне розпізнавання емоцій та його актуальність**

Розпізнавання емоцій стало ключовою сферою в людинно-комп’ютерній взаємодії, з застосуваннями в охороні здоров’я, безпеці, маркетингу та обслуговуванні клієнтів. Традиційні методи розпізнавання емоцій орієнтуються на одномодальні входи, такі як вирази обличчя (FER)[4] або розпізнавання емоцій у мовленні (SER)[5]. Однак мультимодальний підхід, що поєднує текстові, візуальні та аудіодані, значно підвищує точність розпізнавання, використовуючи комплементарність різних модальностей.

Обробка мультимодальних даних вимагає великих та різноманітних наборів даних для охоплення варіацій у вираженні емоцій. Одним з найпоширеніших наборів даних є CMU-MOSEI, який містить понад 65 годин анотованого відеоконтенту, що робить його ідеальним для навчання та валідації мультимодальних нейронних мереж.

**2. Архітектури нейронних мереж для мультимодального розпізнавання емоцій**

Це дослідження розглядає кілька архітектур для обробки мультимодальних даних:

* **CNN для розпізнавання виразів обличчя**: згорткові нейронні мережі (CNN) витягують просторові ознаки з зображень обличчя для ідентифікації емоцій.
* **LSTM для обробки мовлення**: довготривала короткочасна пам’ять (LSTM) аналізує часові залежності в аудіосигналах для розпізнавання емоцій на основі мовлення.
* **Трансформери для аналізу тексту**: трансформери, такі як моделі T5, обробляють текстові входи для аналізу сентименту та емоційного контексту.
* **Моделі мультимодального злиття**: порівнюються різні стратегії злиття, включаючи раннє злиття (поєднання ознак на вході) та пізнє злиття (об’єднання прогнозів окремих моделей), щоб визначити найефективніший підхід до інтеграції.

На рисунку 1 представлено схему запропонованої архітектури мультимодальної моделі.



Рис. 1. Схема мультимодальної моделі

**3. Попередня обробка даних та навчання моделі**

Попередня обробка забезпечує стандартизацію вхідних даних для оптимальної роботи моделі. Це включає:

* **Обробку зображень обличчя**: кадри витягуються з відео, конвертуються у відтінки сірого, а риси обличчя визначаються за допомогою OpenCV.
* **Виділення ознак аудіо**: мовні сигнали знижуються до стандартної частоти дискретизації та перетворюються у мел-спектрограми.
* **Токенізацію тексту**: речення токенізуються за допомогою попередньо навчених токенізаторів T5 для полегшення текстової класифікації емоцій.

**4. Результати та обговорення**

Розроблена мультимодальна модель досягла значного покращення продуктивності порівняно з одномодальними підходами, де результат був в межах 21-24%. Детальний аналіз показав, що точність розпізнавання в реальних умовах досягла 51%, що є значним досягненням. На рисунку 2 показано графік точності навчання та тестування, що демонструє поступове покращення результатів під час навчання.



Рис. 2. Графік точності навчання моделі

Аналіз помилок моделі (рис. 3) свідчить, що система найкраще розпізнає позитивні емоції (98%), тоді як нейтральні емоції мають найнижчий показник точності (67%). Це пояснюється суб'єктивністю нейтрального стану, що підтверджується результатами інших досліджень у цій галузі.



Рис. 3 Аналіз помилок моделі

**5. Перспективи та виклики**

Одним із ключових напрямків подальших досліджень є оптимізація алгоритмів, що дозволить покращити швидкість обробки відеопотоків у режимі реального часу. Крім того, інтеграція WebRTC дозволить значно знизити затримку передавання даних.

Наступним етапом буде розширення навчального набору даних шляхом використання нових джерел відеоконтенту, що містять різноманітні емоційні вирази в реальних умовах.

**Висновки**

Запропонована мультимодальна система розпізнавання емоцій демонструє значні переваги над одномодовими підходами, забезпечуючи точніше та надійніше визначення психоемоційного стану користувачів. Подальші дослідження будуть спрямовані на вдосконалення алгоритмів, розширення наборів даних, оптимізацію роботи системи в умовах реального часу та покращення мультимодального злиття.

**Література**

1. Mayer, J. D., Roberts, R. D., & Barsade, S. G. (2008). Human abilities: Emotional intelligence. Annual Review of Psychology, 59(1), 507–536. <https://doi.org/10.1146/annurev.psych.59.103006.093646>
2. Li, Z. (2017). Facial expression recognition using convolutional neural networks [Video]. YouTube. Retrieved from <https://www.youtube.com/watch?v=zXntbSlqjds>
3. Bui, H., & Tien, L. (2021). Facial expression recognition with CNN-LSTM. Proceedings of the Conference, 549–560. https://doi.org/10.1007/978-981-15-7527-3\_52
4. Song, Z. (2021). Facial expression emotion recognition model integrating philosophy and machine learning theory. Frontiers in Psychology, 12, 759485. <https://doi.org/10.3389/fpsyg.2021.759485>
5. Trends, M. (2020, July 25). Speech emotion recognition (SER) through machine learning. Analytics Insight. Retrieved from https://www.analyticsinsight.net/speech-emotion-recognition-ser-through-machine-learning/