

АНАЛІЗ СТРУКТУРИ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ ДЛЯ РОЗПІЗНАВАННЯ ЕМОЦІЙНОЇ ТОНАЛЬНОСТІ АУДИОЗАПИСІВ

Вінницький національний технічний університет

Анотація

У роботі проведено огляд методів машинного навчання для розпізнавання емоційної тональності аудіозаписів. Показано, що традиційні алгоритми (k -NN, лінійні моделі, дерева рішень) обмежені у врахуванні часових закономірностей, адже перетворюють тривимірні спектрально-часові ознаки у двовимірні представлення. Розглянуто можливості штучних нейронних мереж, зокрема RNN, LSTM, механізми уваги та Conv1D, які зберігають часо-частотну структуру сигналу й виявляють локальні та глобальні патерни. Підкреслено важливість вибору архітектурних блоків для забезпечення точності класифікації і збереження часових залежностей у аудіоданих.

Ключові слова: емоційна тональність, аудіозапис, класифікація, машинне навчання, довга короткочасна пам'ять, багатоголова увага, згортковий шар.

Abstract

This paper reviews machine learning methods for recognizing the emotional tone of audio recordings. It is shown that traditional algorithms (k -NN, linear models, decision trees) are limited in taking into account temporal patterns, since they convert three-dimensional spectral and temporal features into two-dimensional representations. The possibilities of artificial neural networks, in particular RNN, LSTM, attention mechanisms, and Conv1D, which preserve the time-frequency structure of the signal and detect local and global patterns, are considered. The importance of choosing architectural blocks to ensure classification accuracy and preservation of time dependencies in audio data is emphasized.

Keywords: emotional tone, audio recording, classification, machine learning, long short-term memory, multi-headed attention, convolutional layer.

Вступ

Машинне навчання відображає сукупність методів та технологій, які дозволяють створювати моделі для вирішення практичних завдань на основі даних. Цей підхід, що належить до ширшої сфери науки про дані, передбачає етап «навчання» моделі на заздалегідь зібраному масиві інформації [1].

Один із підходів машинного навчання — навчання з учителем — полягає в тому, що модель отримує на вході приклади, кожен із яких містить набір ознак і відповідну мітку чи значення цільової змінної. Завдання алгоритму — побудувати внутрішнє відображення між вхідними даними й мітками, щоб потім точно передбачати результат для нових, раніше не бачених прикладів. У контексті аналізу емоційної тональності аудіозаписів це означає, що кожен фрагмент звуку вже маркований відповідною емоційною категорією, і модель навчається розпізнавати закономірності, які дають змогу правильно впізнати емоційний стан мовця.

У межах навчання з учителем найчастіше виділяють дві категорії завдань — класифікацію та регресію. Завдання класифікації означає, що для кожного входу потрібно віднести приклад до однієї з фіксованих категорій. Натомість у задачах регресії потрібно передбачити неперервне значення. В аналізі емоційної тональності аудіозаписів головним чином мова йде про класифікацію, оскільки потрібно вибрати одну з кількох заздалегідь визначених емоцій.

Усі традиційні методи машинного навчання працюють з двовимірними представленнями даних, де одна вісь — це індекс прикладу, а друга — набір ознак. Однак в аналізі аудіоданих часто відбувається робота з тривимірними структурами даних. У підсумку жоден зі стандартних підходів не зберігає повноцінну часову структуру, яка надзвичайно важлива для розпізнавання емоцій. Адже саме динаміка інтонації, швидкість змін тембру, чергування гучності тощо можуть дати модельній системі важливі підказки про емоційний стан мовця.

Саме тому для задачі впізнавання емоційної тональності аудіозаписів дедалі частіше звертаються до штучних нейронних мереж, які здатні працювати з багатовимірними вхідними даними в їхньому

початковому вигляді. Базовим елементом такої мережі є штучний нейрон, що імітує спрощену модель біологічного нейрона. На вході він отримує набір сигналів (ознаки), кожен із яких має свою вагу. Далі нейрон обчислює зважену суму цих сигналів, а потім передає результат через нелінійну функцію активації, яка вирішує, наскільки сильним буде вихідний сигнал. Завдяки такій структурі можна накопичувати інформацію й пріоритизувати важливі патерни з багатовимірного сигналу [2].

Результати дослідження

Штучна нейронна мережа складається з кількох шарів таких нейронів: вхідного, одного або кількох прихованих та вихідного. Данні проходять через мережу поступово — спочатку крізь перший прихований шар, потім другий, і так далі, а на виході отримуємо вектори, які інтерпретуються як ймовірності належності до різних емоційних класів. У процесі навчання ваги кожного зв'язку між нейронами оновлюються так, щоб мережа давала якомога точніші прогнози.

У випадку обробки послідовних аудіоданих важливо враховувати їхню часову динаміку. Повнозв'язні (Dense) шари не здатні «запам'ятовувати» послідовність у прямому вигляді, бо кожен вихідний нейрон отримує незалежне представлення всіх входів, й інформація про порядок фреймів моментально втрачається. Тому для покращення результатів застосовують спеціальні структури.

Одним із прикладів таких структур є рекурентні нейронні мережі, в яких кожен елемент обробляє один часовий крок, а внутрішній стан (прихований стан) передається далі, що дає можливість враховувати контекст попередніх фреймів. У теорії це підходить для послідовностей, але на практиці при довгих аудіо виникають проблеми: через затухання або вибух градієнтів мережа не може ефективно «пам'ятати» події, що траплялися задовго до поточного моменту. Як наслідок, модель може «забути» важливі патерни, які впливали б на класифікацію емоції.

Щоб обійти проблему «забування», введено спеціалізовані блоки з довгою короткочасною пам'яттю. Кожний такий елемент має додаткові механізми: вектори «забування» й «запам'ятовування», які самостійно вирішують, яке інформаційне навантаження важливо залишити, а яке — ігнорувати [3]. Завдяки цьому LSTM-мережі можуть працювати з довшими послідовностями і зберігати ключові закономірності, необхідні для правильної класифікації емоцій (рис. 1).

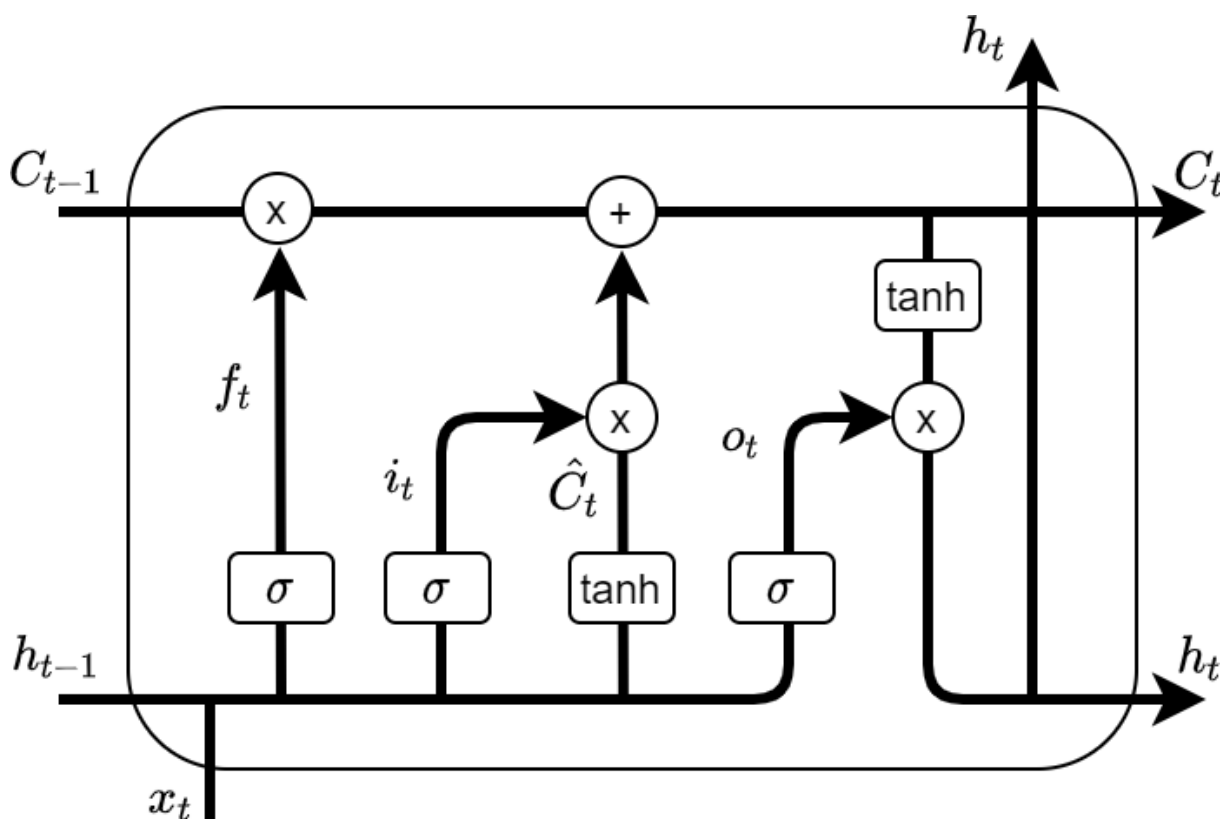
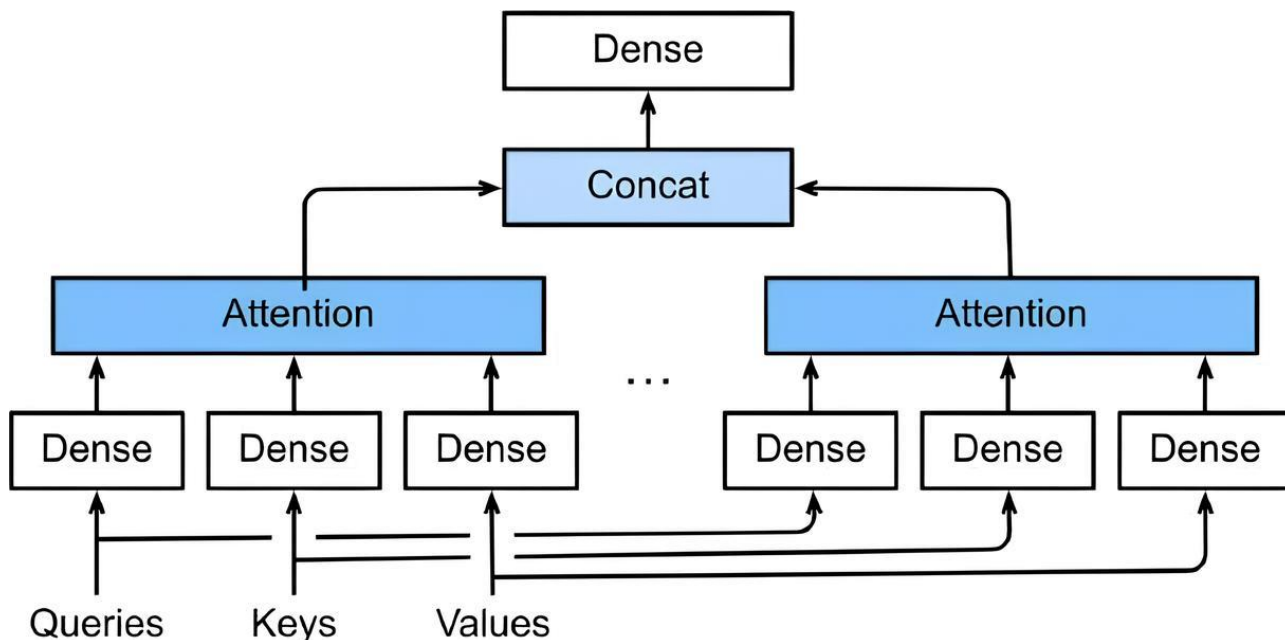


Рис. 1. Спрощена модель елемента з довгою короткочасною пам'яттю

Одним із головних обмежень рекурентних мереж є те, що вони обробляють дані послідовно, і кожний крок залежить від попереднього. Блоки з увагою дозволяють моделі «дивитися» на всю послідовність одразу, виділяючи найважливіші частини для поточного етапу обробки. Таким чином,

замість того, щоб поступово «спускатися» від початку фрагмента до кінця, мережа обчислює ваги уваги для кожного фрейма й вирішує, які з них найбільш інформативні.



Ще одним важливим класом архітектур для аудіоданих є згорткові нейронні мережі (CNN). У традиційних задачах обробки зображень використовують двовимірні фільтри, утворюючи карти ознак, які виявляють контури, текстури та інші просторові патерни. Але для звуку найчастіше застосовують одновимірні згортки, які «ковзають» по часовій вісі, аналізуючи одновимірний сигнал або його попередньо видобуті спектральні репрезентації (наприклад, мел-спектрограми). Кожен фільтр навчиться знаходити певні характерні патерни, наприклад, швидкі стрибки амплітуди, характерні для окремих емоційних сплесків [5] (рис. 3).

Рис. 3. Принцип роботи одновимірного фільтра згорткового шару

У цій роботі було проведено аналіз структури моделей машинного навчання для розпізнавання емоційної тональності аудіозаписів. Показано, що класичні алгоритми (k-NN, лінійні моделі, дерева рішень і їхні ансамблі) мають обмежену здатність враховувати часову динаміку аудіоданих, оскільки змушені трансформувати тривимірні спектрально-часові ознаки у двовимірні вектори або матриці.

Виділено ключові архітектури для обробки послідовних аудіоданих: рекурентні мережі (RNN), їхні модифікації з довгою короткочасною пам'яттю (LSTM) та механізми багатоголової уваги. Показано, що такі блоки дозволяють зберегти й ефективніше використовувати часові залежності у фічах, необхідні для точного класифікування емоцій. Також окремою групою розглянуто згорткові нейронні мережі (Conv1D), які «зчитують» локальні спектрально-часові патерни.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Мокін В. Б., Дратований М. В. Наука про дані: машинне навчання та інтелектуальний аналіз даних : електронний навчальний посібник комбінованого (локального та мережевого) використання. Вінниця : ВНТУ, 2024. 258 с.
2. Мілян Н. Аналіз методів машинного навчання з вчителем. *Міжнародна студентська науково-технічна конференція "Природничі та гуманітарні науки. Актуальні питання"* : матеріали конф., ТНТУ імені Івана Пулюя. URL: https://elartu.tntu.edu.ua/bitstream/lib/25035/2/MSNK_2018v1_Milian_N-Analysis_of_supervised_machine_51-52.pdf
3. Лосенко А. В., Козачко О. М., Варчук І. В. Нейромережевий ансамбль для прогнозування часових рядів на основі Prophet та LSTM. *Наукові праці Вінницького національного технічного університету*. 2024. №4. URL: <https://doi.org/10.31649/2307-5376-2024-4-49-57>
4. Chen Y., Pu H., Qu Y. An analysis of attention mechanisms and its variance in transformer. *Applied and Computational Engineering*. 2024. №47. С. 164–176. URL: <https://doi.org/10.54254/2755-2721/47/20241291>
5. Bhatt M., Sharma A., Singh A. A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*. 2024. №57. URL: <https://doi.org/10.1007/s10462-024-10721-6>

Рудзевич Олександр Володимирович — студент групи СА-21б, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, м. Вінниця, e-mail: sasha.rrrr.1@gmail.com

Жуков Сергій Олександрович — к.т.н., доцент кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, e-mail: sazhukov@gmail.com

Rudzevych Olexandr V. — student of group SA-21b, Faculty of Intelligent Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail: sasha.rrrr.1@gmail.com

Zhukov Serhii O. — Ph.D., Assistant Professor of the Department of Systems Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: sazhukov@gmail.com