УДК 004.8 DOI: 10.31673/2786-8362.2025.021372

Бай Я.В.; Катков Ю.І., д.т.н.

ДОСЛІДЖЕННЯ ГІБРИДНОЇ АРХІТЕКТУРИ CNN З RESIDUAL SHUFFLE-EXCHANGE ДЛЯ РОЗПІЗНАВАННЯ НОТ З АУДІОДАНИХ

Bai Ya.V., Katkov Yu.I. Research on hybrid CNN architecture with Residual Shuffle-Exchange for note recognition from audio data. The article discusses modern neural network approaches to automatic note recognition from audio files. Algorithms based on fast Fourier transform, convolutional neural networks (CNN), and residual shuffle-exchange (RSE) networks are analyzed. The advantages of deep learning in conditions of background noise and variable note performance are identified. Ways to improve recognition accuracy and prospects for application in music technologies are proposed.

Keywords: Note recognition, audio analysis, neural networks, convolutional neural network, RSE network, music transcription, deep learning, sound processing, machine learning, artificial intelligence

Бай Я.В., Катков Ю.І. Дослідження гібридної архітектури CNN з Residual Shuffle-Exchange для розпізнавання нот з аудіоданих. У статті розглядаються сучасні нейромережеві підходи до автоматичного розпізнавання нот з аудіофайлів. Проаналізовано алгоритми на основі швидкого перетворення Фур'є, згорткових нейронних мереж (CNN) та мереж залишкових перестановок-обмінів (RSE). Визначено переваги глибокого навчання в умовах фонових шумів і варіативного виконання нот. Запропоновано шляхи підвищення точності розпізнавання та перспективи застосування в музичних технологіях.

Ключові слова: Розпізнавання нот, аудіоаналіз, нейронні мережі, згорткова нейронна мережа, RSE-мережа, музична транскрипція, глибоке навчання, обробка звуку, машинне навчання, штучний інтелект

Вступ

Від давніх ритмів до сучасних цифрових композицій музика супроводжує людство протягом тисячоліть. Вона не потребує перекладу - мелодія, створена в одному куточку світу, може торкнутися серця слухача на іншому континенті. Музика здатна передавати те, що неможливо висловити словами: радість, сум, надію, ностальгію. Однак коли ми намагаємося навчити машину "розуміти" музику так, як це робить людина, ми стикаємося з фундаментальними викликами, які розкривають справжню складність цього мистецтва.

Розпізнавання нот – це процес ідентифікації окремих нот з аудіофайлу, який на перший погляд може здаватися простим. Людське вухо легко розрізняє ноти, проте для комп'ютера це надзвичайно складна задача. Причина полягає в принциповій різниці між нотною партитурою та аудіосигналом: якщо на папері кожна нота чітко визначена своїм положенням на нотоносці, то в реальному звучанні ми маємо справу з безперервною хвилею, в якій ноти накладаються одна на одну, змішуються з гармоніками та спотворюються різними акустичними ефектами.

Найбільшу складність становить поліфонічність музики [1]. Коли декілька інструментів або голосів звучать одночасно, їхні частоти змішуються в єдиний складний сигнал. Виділити окремі ноти з цієї суміші – все одно що намагатися розділити воду, каву та молоко після того, як їх змішали в одній чашці. Якщо традиційні алгоритми добре справляються з монофонічною музикою (одна нота в момент часу), то при роботі з поліфонією вони зазнають невдачі. До цього додається проблема тембрального різноманіття інструментів [2]. Нота "ля" першої октави має частоту 440 Гц незалежно від інструменту, але звучить кардинально по-різному на флейті, гітарі чи фортепіано. Це пояснюється наявністю обертонів – додаткових частот, які супроводжують основну. Кожен інструмент має унікальний спектральний відбиток, тому система розпізнавання повинна вміти ідентифікувати ноту, ігноруючи тембральні відмінності. Ситуацію ускладнюють і динамічні зміни з артикуляцією. Музиканти не грають ноти як ідеальні прямокутні імпульси – ці ноти можуть звучати зовсім по-різному в залежності від нюансів виконання. Початок ноти може містити перехідні процеси з широким спектром частот, які важко інтерпретувати. Більше того, гучність ноти постійно змінюється, що впливає на видимість її гармонік у спектрі.

У реальних умовах задача стає ще складнішою через фоновий шум та акустичне середовище [3]. Реверберація в приміщенні створює відлуння, яке змазує межі між нотами. Шум від аудиторії, дихання музиканта, скрип педалей фортепіано – все це додає непотрібні компоненти до сигналу. Стиснення аудіо з втратами (MP3, AAC) також вносить артефакти, які можуть маскувати або спотворювати справжні ноти. Окремою проблемою є темпоральна невизначеність. Точно визначити, коли саме починається і закінчується нота, часто неможливо. У швидких музичних пасажах ноти можуть бути дуже короткими, і традиційні методи аналізу з використанням віконного перетворення Фур'є стикаються з компромісом між частотною та часовою роздільною здатністю. Додаткові виклики створює музичний контекст та особливості нотації. Деякі музичні техніки, такі як бенди на гітарі, портаменто у співу або мікротональні відхилення, важко представити в стандартній нотації. Система повинна вирішити, чи слід інтерпретувати поступові зміни висоти як окремі ноти чи як один плавний перехід.

Нарешті, розвиток цієї галузі гальмує відсутність стандартизованих наборів даних, що ускладнює навчання та оцінку систем розпізнавання. Різні дослідники використовують різні набори даних, критерії оцінки та методології, що робить порівняння результатів проблематичним.

З появою глибокого навчання та нейронних мереж з'явилися нові можливості для вирішення цих викликів. Сучасні архітектури здатні вчитися складним залежностям безпосередньо з даних, проте навіть вони не досягли людської точності в складних випадках. Розуміння цих фундаментальних проблем залишається ключем до розробки більш ефективних систем автоматичного розпізнавання нот.

Аналіз останніх досліджень. Автоматична музична транскрипція (Automatic Music Transcription, AMT) є однією з найскладніших та найцікавіших задач цифрової обробки аудіосигналів. Ця задача полягає у перетворенні музичного запису в символьне представлення – нотний текст або MIDI-файл, що містить інформацію про висоту, тривалість та динаміку кожної ноти. Протягом останніх десятиліть підходи до вирішення цієї проблеми пройшли значний еволюційний шлях – від класичних методів обробки сигналів, які базувалися на математичних моделях та експертних знаннях, до сучасних архітектур глибинного навчання, здатних автоматично витягувати складні ознаки з аудіоданих [4].

Складність AMT обумовлена необхідністю одночасного вирішення декількох підзадач: визначення висоти тону (pitch detection), виявлення моментів початку та закінчення нот (onset/offset detection), розділення джерел звуку у поліфонічних записах (source separation), а також врахування музичної експресії та тембральних характеристик інструментів. При цьому алгоритми повинні бути робастними до шумів, реверберації, варіацій темпу та інших факторів, характерних для реальних музичних записів.

Традиційні методи розпізнавання нот базуються на класичних підходах цифрової обробки сигналів та математичному моделюванні властивостей звуку. Ці алгоритми розроблялися з урахуванням фізичних принципів звукоутворення та психоакустичних закономірностей сприйняття музики людиною. Основною перевагою таких методів є їхня інтерпретованість – кожен крок обробки має чітке математичне обґрунтування, а результати можна аналізувати та коригувати на основі експертних знань. Крім того, традиційні алгоритми зазвичай мають низькі обчислювальні вимоги та не потребують великих обсягів навчальних даних, що робить їх придатними для роботи в реальному часі на обладнанні з обмеженими ресурсами.

Традиційні підходи можна умовно поділити на три великі категорії залежно від того, в якій області відбувається аналіз сигналу: *методи часової області* працюють безпосередньо з формою хвилі аудіосигналу, *методи частотної області* аналізують спектральний склад звуку, а *класичні підходи до поліфонії* використовують комбінацію різних технік для розділення одночасно звучащих нот.

Революція глибинного навчання, що розпочалася в 2010-х роках, кардинально змінила підходи до музичної транскрипції. На відміну від традиційних методів, які потребували

ручного проектування ознак та експертних знань про структуру музичних сигналів, нейромережеві підходи здатні автоматично навчатися ієрархічним представленням даних безпосередньо з великих обсягів прикладів. Це дозволило моделям виявляти складні закономірності, які важко формалізувати аналітично, та адаптуватися до різноманітності музичних стилів, інструментів та умов запису.

Методи глибинного навчання для музичної транскрипції можна класифікувати за типом архітектури нейронної мережі. Згорткові нейронні мережі (CNN) ефективно обробляють часово-частотні представлення звуку, такі як спектрограми, виявляючи локальні закономірності гармонічних структур. Рекурентні мережі (RNN/LSTM) спеціалізуються на моделюванні темпоральних залежностей, відстежуючи розгортання музичних композицій у часі. Трансформери використовують механізм уваги для виявлення довгострокових залежностей та музичної структури. При цьому генеративні моделі пропонують принципово новий погляд на транскрипцію як на задачу генерації нотного тексту з аудіо. Розглянемо кожну з цих груп детальніше.

Згорткові нейронні мережі природно підходять для аналізу музичних спектрограм завдяки своїй здатності виявляти локальні закономірності та інваріантність до зсувів. Даний тип нейронних мереж традиційно використовувався для розв'язання задачі розпізнавання зображень. Спектрограма, по суті, є зображенням, де горизонтальна вісь представляє час, вертикальна – частоту, а інтенсивність пікселів – енергію сигналу. Згорткові шари послідовно застосовують набори фільтрів до цього "зображення", виявляючи спочатку прості ознаки (краї, текстури), а на вищих рівнях – більш абстрактні паттерни, такі як гармонічні структури або тембральні характеристики інструментів.

Базові CNN-архітектури для музичної транскрипції зазвичай складаються з кількох згорткових блоків, кожен з яких включає згортковий шар, функцію активації (часто це функція ReLU) та операцію пулінгу для зменшення розмірності. Для монофонічної транскрипції вихідний шар мережі є класифікатором, який визначає висоту тону в кожному часовому кадрі. Глибокі архітектури типу VGG або ResNet, адаптовані для аудіо даних, демонструють точність 95-98% на чистих студійних записах окремих інструментів.

Для поліфонічної транскрипції CNN працюють як multi-label класифікатори, де кожен з виходів (зазвичай 88, по кількості клавіш фортепіано) незалежно передбачає наявність відповідної ноти. Особливо цікавою є модель **Deep Salience** [5], яка використовує паралельні CNN-гілки для обробки різних частотних діапазонів. Кожна гілка спеціалізується на певному регістрі, що дозволяє краще розділяти ноти, які близькі за висотою. Виходи всіх гілок потім об'єднуються для формування фінального передбачення. Така архітектура досягає точності 70-75% для поліфонічної транскрипції, що є значним покращенням порівняно з традиційними методами.

Музика є суттєво темпоральним феноменом – розуміння музичного контексту вимагає врахування не лише поточного моменту, але й того, що звучало раніше і що очікується далі. Рекурентні нейронні мережі спеціально розроблені для обробки послідовних даних, підтримуючи внутрішній "стан пам'яті", який оновлюється на кожному кроці і несе інформацію про попередній контекст.

Bidirectional LSTM (Long Short-Term Memory) [6] виявилися особливо ефективними для музичної транскрипції. На відміну від простих RNN, LSTM мають спеціальну архітектуру "воріт", яка дозволяє селективно запам'ятовувати важливу інформацію та забувати нерелевантну, вирішуючи проблему зникаючого градієнта при навчанні на довгих послідовностях. Двонаправлений варіант обробляє послідовність одночасно в прямому та зворотному напрямках, що допомагає краще локалізувати момент закінчення поточної ноти та виправити помилки визначення початку. Часто LSTM комбінують з умовними випадковими полями (**LSTM-CRF** архітектури), які додають структурні обмеження на послідовність передбачень. Така комбінація покращує узгодженість транскрипції з музичними правилами та підвищує загальну точність системи до 72-78% для поліфонічних записів.

Transformer-архітектури, які революціонізували обробку природної мови, знайшли своє застосування і в музичній транскрипції. Ключовою ідеєю трансформерів є механізм self-

attention, який дозволяє моделі безпосередньо звертатися до будь-якого моменту вхідної послідовності, незалежно від відстані, на відміну від RNN, які послідовно обробляють інформацію крок за кроком. Це дає можливість захоплювати довгострокові залежності та музичну структуру на рівні фраз, мотивів та цілих секцій композиції.

Проте класичний механізм self-attention має квадратичну обчислювальну складність відносно довжини послідовності, що робить його непрактичним для довгих музичних записів. **Residual Shuffle-Exchange Networks (RSE)** [7] пропонують елегантне рішення цієї проблеми, досягаючи лінійної складності. Ця архітектура базується на ідеї shuffle-exchange мереж з теорії паралельних обчислень, адаптованих для нейронних мереж через residual з'єднання, активацію GELU та Layer Normalization. RSE дозволяє ефективно обробляти послідовності довжиною до 2 мільйонів елементів, що відповідає годинам аудіо, працюючи безпосередньо з формою хвилі без попереднього обчислення спектрограм. На датасеті MusicNet ця модель досягла state-of-the-art результатів з точністю 78-82% для поліфонічної транскрипції, будучи при цьому компактною за кількістю параметрів та швидкою в навчанні.

Останнім часом з'явився новий погляд на транскрипцію через призму генеративних моделей. Замість того, щоб напряму передбачати ноти, ці підходи навчаються моделювати розподіл можливих транскрипцій для даного аудіо та генерувати найбільш правдоподібний варіант.

Variational Autoencoders навчаються стискати музичні дані в зменшене латентне представлення, з якого потім можна реконструювати вихідну інформацію. При цьому **diffusion моделі** розглядають транскрипцію як задачу поступового "очищення" від шуму. Такий підхід дозволяє генерувати високоякісні транскрипції навіть з сильно зашумлених або спотворених записів, досягаючи точності 78-83% для поліфонічної музики. Однак ці моделі вимагають багаторазового проходження через мережу під час інференсу, що робить їх значно повільнішими за інші підходи.

Постановка завдання. Отже, перспективним компромісом для розпізнавання нот з аудіоданих є гібридні підходи, що поєднують сильні сторони обох парадигм. Наприклад, використання CQT як вхідного представлення для CNN дозволяє використати переваги логарифмічної частотної шкали, природної для музики, разом з потужністю глибинного навчання. Або застосування NMF для попереднього розділення джерел з подальшою нейромережевою обробкою кожного джерела окремо.

Метою роботи є розробити та дослідити ефективність гібридної нейронної мережі на основі згорткових нейронних мереж (CNN) та архітектури Residual Shuffle Exchange (RSE) для автоматичного розпізнавання музичних нот з аудіозаписів.

Виклад основного матеріалу дослідження

При проектуванні гібридної архітектури CNN+RSE виникає ключове питання: яким чином оптимально поєднати ці два типи мереж для максимального використання їх переваг та мінімізації недоліків? Існує декілька фундаментальних стратегій архітектурного дизайну, кожна з яких має різні компромісі між продуктивністю, обчислювальною складністю та складністю реалізації.

Послідовна композиція є найбільш інтуїтивним підходом, де виходи одного типу мережі служать входами для іншого. Ця стратегія природно відповідає ієрархічній обробці інформації: низькорівневі ознаки екстрагуються першим компонентом, а високорівневі залежності моделюються другим. Питання полягає в тому, який порядок є оптимальним: CNN→RSE чи RSE→CNN, та скільки блоків кожного типу необхідно використовувати.

Паралельна обробка дозволяє різним типам мереж працювати незалежно над одними й тими самими вхідними даними, після чого їх виходи об'єднуються. Ця стратегія забезпечує максимальне збереження інформації, оскільки жоден тип обробки не залежить від іншого та не обмежується його помилками. Однак паралельні архітектури вимагають більше обчислювальних ресурсів та пам'яті, а також потребують ефективних механізмів злиття (fusion) інформації з різних гілок.

Чергуюча архітектура передбачає багаторазове чергування різних типів обробки протягом мережі. Цей підхід дозволяє моделі уточнити представлення на кожному рівні абстракції, поєднуючи просторову та темпоральну інформацію ітеративно. Хоча чергуючі архітектури потенційно найбільш виразні, вони також є найскладнішими в проектуванні та навчанні через необхідність узгодження форматів даних та ризик деградації градієнтів у глибоких мережах.

Вибір оптимальної стратегії поєднання CNN та RSE залежить від багатьох факторів: характеристик даних (монофонічна чи поліфонічна музика, тривалість фрагментів), доступних обчислювальних ресурсів, вимог до латентності при інференсі, та специфічних особливостей задачі транскрипції.

Послідовний підхід. Найбільш інтуїтивним та широко використовуваним способом поєднання CNN та RSE є послідовна архітектура, де згорткові шари виконують роль екстракторів ознак, а RSE-блоки працюють як механізм темпорального моделювання. Цей підхід природно відповідає ієрархічній природі обробки музичного сигналу, де спочатку необхідно виділити базові акустичні характеристики, а потім встановити зв'язки між ними в часі.

У цій архітектурі вхідна спектрограма спочатку проходить через кілька згорткових блоків, які поступово збільшують глибину представлення та зменшують просторову розмірність. Перші згорткові шари навчаються розпізнавати прості частотні патерни, такі як гармоніки окремих нот, їх основні частоти та обертони. Наступні шари комбінують ці прості ознаки в більш складні структури, здатні розпізнавати специфічні тембральні характеристики інструментів та одночасно звучачі акорди. Використання двовимірних згорток критично важливе на цьому етапі, оскільки музичні патерни мають чітку структуру як у частотній, так і в часовій осях.

Після CNN-блоків відбувається критичний етап трансформації представлення. Двовимірні карти ознак необхідно перетворити в послідовність для подальшої обробки RSE-блоками. Це можна здійснити через операцію reshape, яка зберігає часову вісь, але згортає частотну вісь та канали в єдиний вектор ознак для усіх часових інтервалів [8]. Важливо на цьому етапі зберегти темпоральну структуру, оскільки саме вона буде ключовою для наступного етапу обробки.

RSE-блоки отримують послідовність векторів ознак і моделюють складні часові залежності між ними. Архітектура Residual Shuffle Exchange особливо ефективна для музичної транскрипції завдяки своїй здатності обробляти довгі послідовності без проблеми зникаючих градієнтів [9]. Shuffle-операції в RSE дозволяють різним частинам представлення обмінюватися інформацією, що критично важливо для розуміння поліфонічної музики, де одночасно звучать декілька нот і їх взаємодія визначає музичний контекст. Залишкові з'єднання забезпечують стабільність навчання та дозволяють інформації з попередніх шарів безпосередньо впливати на прийняття фінальних рішень [10].

Основною перевагою послідовного підходу є його концептуальна ясність та ефективність навчання. CNN-блоки швидко навчаються розпізнавати стабільні акустичні патерни, створюючи компактне представлення вхідного сигналу. RSE-блоки працюють з цим вже очищеним та структурованим представленням, що значно спрощує їхню задачу та прискорює збіжність. Крім того, така архітектура добре масштабується та дозволяє легко контролювати баланс між просторовою та темпоральною обробкою шляхом варіювання кількості CNN та RSE блоків.

Однак послідовний підхід має свої обмеження. Він передбачає, що вся просторова обробка повинна бути завершена до початку темпорального моделювання, що може призвести до втрати деяких важливих деталей. Наприклад, CNN може згорнути інформацію про тонкі частотні відмінності, які могли б бути важливими для розрізнення близьких за висотою нот. RSE-блоки не мають можливості повернутися до вихідного спектрального представлення і можуть працювати лише з тим, що їм передали згорткові шари. Схему послідовного підходу зображено на рисунку 1.

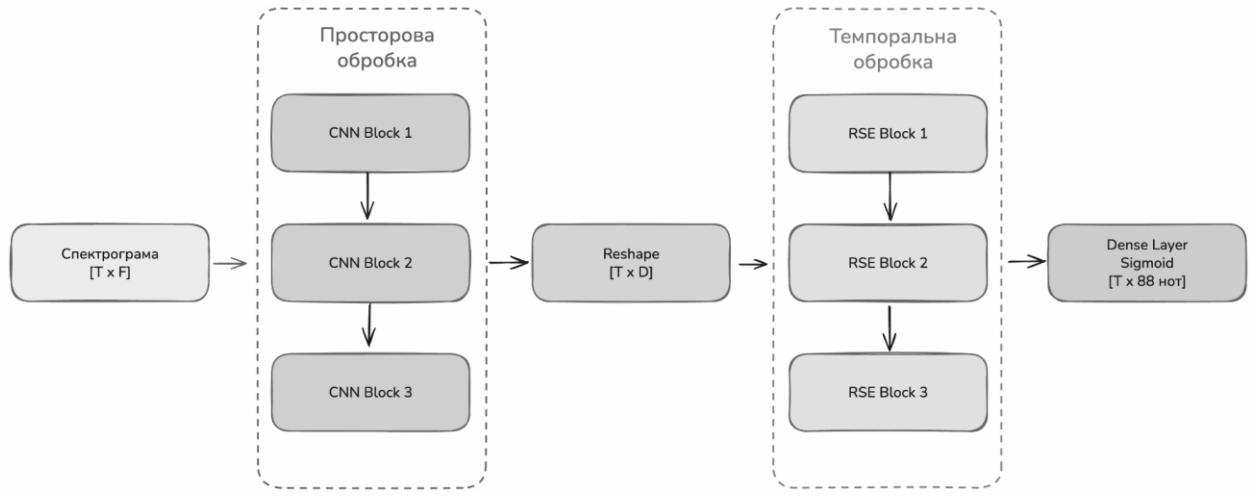

Рис. 1. Схема послідовного підходу

Паралельний підхід. Паралельна архітектура представляє принципово інший спосіб поєднання CNN та RSE, де обидва типи обробки відбуваються одночасно над різними формами вхідних даних. У цьому підході вхідна спектрограма подається одночасно в дві окремі гілки обробки [11]. Перша гілка містить згорткові блоки, які виконують просторову екстракцію ознак так само, як у послідовному підході. Друга гілка працює безпосередньо з часовою послідовністю спектральних векторів, використовуючи RSE-блоки для моделювання темпоральних залежностей без попередньої згортки.

Ключова ідея паралельного підходу полягає в тому, що різні аспекти музичного сигналу можуть вимагати різних стратегій обробки [12]. CNN-гілка фокусується на виявленні стійких спектральних патернів, таких як гармонічні структури нот, їх формантні характеристики та тембральні особливості інструментів. Ця гілка ефективно працює з просторовими кореляціями в спектрограмі, виділяючи інваріантні до часу ознаки. RSE-гілка, навпаки, концентрується на динамічних аспектах музики, таких як ритмічні патерни, переходи між нотами, атаки та затухання звуку, а також загальний темпоральний контекст музичної фрази [13].

Після обробки в окремих гілках представлення об'єднуються через механізм конкатенації або більш складні методи злиття, такі як attention-based fusion [14]. Конкатенація є найпростішим варіантом, де вектори ознак з обох гілок просто з'єднуються разом, створюючи багатший опис кожного часового моменту. Механізм уваги дозволяє моделі динамічно визначати, яка гілка більш релевантна для конкретного часового кроку або частотного регіону, що може бути особливо корисним для складної поліфонічної музики.

Основною перевагою паралельного підходу є його здатність зберігати всю інформацію з обох шляхів обробки. Якщо CNN-гілка втрачає певні темпоральні деталі через pooling операції, RSE-гілка все ще має доступ до повної часової роздільності вхідного сигналу. Аналогічно, якщо RSE-гілка не може ефективно виділити тонкі спектральні патерни, CNN-гілка компенсує цей недолік своєю спеціалізованою обробкою. Така архітектура також природно піддається інтерпретації, оскільки можна окремо проаналізувати внесок кожної гілки в фінальне рішення.

Проте паралельний підхід вимагає значно більше обчислювальних ресурсів порівняно з послідовним. Обидві гілки повинні обробляти повнорозмірні вхідні дані, що подвоює витрати пам'яті та обчислень [15]. Крім того, навчання паралельної архітектури може бути складнішим, оскільки необхідно збалансувати навчання обох гілок та забезпечити, щоб вони вивчали комплементарні, а не дублюючі представлення. Часто виникає проблема, коли одна гілка домінує над іншою, і модель фактично ігнорує вихід однієї з гілок. Схему паралельного підходу зображено на рисунку 2.

Чергуючий підхід. Третій варіант архітектури передбачає чергування CNN та RSE блоків протягом всієї мережі. У цьому підході вхідна спектрограма спочатку обробляється

одним або кількома згортковими шарами, потім послідовність передається до RSE-блоку, після чого знову застосовуються згортки, і цей патерн повторюється кілька разів. Така організація дозволяє моделі багаторазово уточнювати представлення, комбінуючи просторову та темпоральну обробку на різних рівнях абстракції.

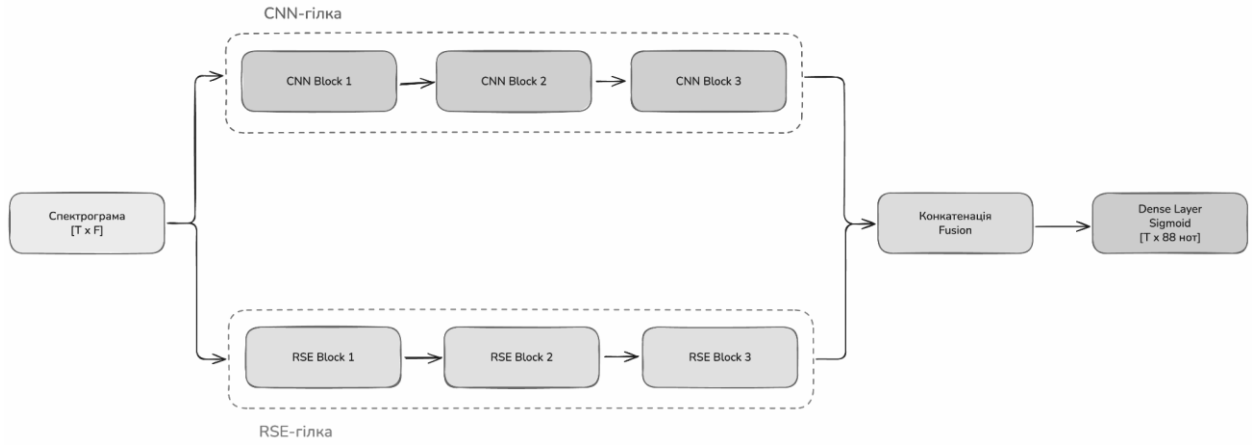

Рис. 2. Схема паралельного підходу

Чергуючий підхід базується на ідеї ієрархічного представлення музики, де на кожному рівні абстракції корисні як просторові, так і темпоральні трансформації. Ранні CNN-блоки виділяють базові акустичні ознаки, перший RSE-блок встановлює початкові часові зв'язки між ними, наступні CNN-блоки комбінують ці контекстуалізовані ознаки в більш складні структури, а наступний RSE-блок моделює взаємодії на вищому рівні семантики. Таке багаторазове чергування дозволяє моделі будувати багатшу внутрішню репрезентацію музичного сигналу.

Важливою технічною деталлю чергуючого підходу є необхідність узгодження форматів даних між CNN та RSE блоками. Після кожного згорткового блоку може знадобитися операція reshape для підготовки даних до RSE, а перед наступним згортковим блоком необхідно відновити двовимірну структуру [16]. Це можна реалізувати через спеціальні адаптивні шари, які навчаються оптимальному способу трансформації представлень. Альтернативно, можна використовувати одновимірні згортки після першого переходу до послідовного формату, що спрощує архітектуру, але втрачає частину переваг двовимірної обробки.

Потенційна перевага чергуючого підходу полягає в його гнучкості та можливості моделювати складні ієрархічні залежності. Музична інформація на різних рівнях абстракції може мати різну природу, і чергування різних типів обробки дозволяє адаптуватися до цієї специфіки. Наприклад, на низькому рівні важливі спектральні патерни окремих нот та їх початки, на середньому рівні стають важливими акордові прогресії та ритмічні структури, а на високому рівні модель може захоплювати музичні фрази та мелодичні контури [17].

Однак чергуючий підхід має найскладнішу архітектуру з усіх розглянутих варіантів. Постійне перетворення між просторовими та послідовними представленнями може призводити до втрати інформації, якщо не використовувати додаткові механізми збереження контексту, такі як skip connections між неоднорідними блоками [18]. Навчання такої мережі також вимагає ретельного налаштування, оскільки глибина мережі швидко зростає, що може призвести до проблем зі збіжністю. Крім того, інтерпретація поведінки моделі стає складнішою, оскільки важко відстежити, як інформація трансформується при кожному переході між типами обробки. Схему чергуючого підходу зображено на рисунку 3.

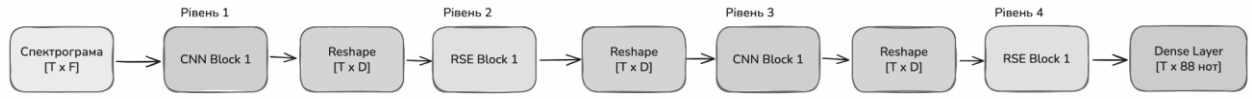

Рис. 3. Схема чергуючого підходу

Вибір оптимального підходу. Вибір між цими трьома підходами залежить від конкретних вимог задачі, доступних обчислювальних ресурсів та характеристик даних. Послідовний підхід є найкращим стартовим варіантом для більшості випадків завдяки своїй простоті, надійності та обчислювальній ефективності. Він добре працює для монофонічної та помірно поліфонічної музики, де чітке розділення просторової та темпоральної обробки є природним. Цей підхід також найлегше налаштувати та відлагодити, що робить його ідеальним для початкових експериментів та встановлення базової лінії продуктивності.

Паралельний підхід варто розглядати для складної поліфонічної музики, де важливо зберегти максимум інформації з вихідного сигналу. Він особливо корисний, коли є достатньо обчислювальних ресурсів і потрібна найвища можлива точність розпізнавання. Паралельна архітектура також має переваги при роботі з різними музичними жанрами в одному наборі даних, оскільки різні гілки можуть спеціалізуватися на різних типах музики.

Чергуючий підхід може показати найкращі результати на особливо складних музичних композиціях із багатим гармонічним змістом та складною ритмічною структурою, але вимагає значних зусиль на етапі проектування та навчання моделі.

Висновки

У цій роботі було представлено комплексне дослідження гібридних архітектур на основі поєднання згорткових нейронних мереж (CNN) та Residual Shuffle-Exchange мереж (RSE) для задачі автоматичного розпізнавання музичних нот з аудіозаписів. Проведений аналіз продемонстрував, що інтеграція цих двох типів архітектур дозволяє ефективно використовувати переваги кожного підходу: просторову екстракцію ознак через згорткові шари та ефективне темпоральне моделювання через RSE-блоки з логарифмічною обчислювальною складністю.

Проведене дослідження продемонструвало значний потенціал гібридних архітектур CNN+RSE для задачі автоматичного розпізнавання музичних нот. Систематичний аналіз трьох стратегій поєднання цих типів мереж надає теоретичну основу для обґрунтованого вибору архітектурних рішень залежно від специфічних вимог та обмежень конкретної задачі. Результати роботи вносять вклад як у теоретичне розуміння принципів побудови гібридних нейромережевих архітектур, так і в практичні аспекти розробки систем музичної транскрипції. Подальші емпіричні дослідження та оптимізація запропонованих підходів.

Результати цього дослідження відкривають декілька перспективних напрямків для подальшої роботи. По-перше, необхідна емпірична валідація запропонованих архітектур на стандартних датасетах, таких як MAPS, MAESTRO та MusicNet, з детальним порівнянням метрик точності, recall, F1-score на рівні frame та note. По-друге, доцільно дослідити гібридні варіанти, що поєднують елементи різних стратегій, наприклад, паралельну обробку на ранніх шарах з послідовною на пізніх. По-третє, перспективним є вивчення механізмів adaptive fusion, які динамічно визначають оптимальний спосіб об'єднання CNN та RSE представлень залежно від характеристик вхідного сигналу.

Додатковим напрямком досліджень є оптимізація гіперпараметрів для кожної архітектури: кількість CNN та RSE блоків, розмір фільтрів, глибина мереж, стратегії regularization. Важливо також дослідити техніки transfer learning та domain adaptation для адаптації моделей, навчених на одних типах музичних інструментів, до інших. Крім того, доцільно вивчити можливості розширення запропонованих архітектур для багатоінструментальної транскрипції з одночасним розпізнаванням тембру та source separation.

Список використаної літератури:

1. Bay M., Ehmann A. F., Downie J. S. Evaluation of multiple-F0 estimation and tracking systems. Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR). 2009. P. 315–320. URL: https://ismir2009.ismir.net/proceedings/PS2-13.pdf.

2. Automatic music transcription: challenges and future directions / E. Benetos et al. Journal of Intelligent Information Systems. 2013. Vol. 41, no. 3. P. 407–434. URL: https://doi.org/10.1007/s10844-013-0258-3.

3. Cogliati A., Temperley D., Duan Z. Transcribing human piano performances into music notation. Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR). 2016. P. 758–764. URL: https://wp.nyu.edu/ismir2016/wp-content/uploads/sites/2294/2016/07/cogliati-transcribing.pdf.

4. Automatic music transcription: challenges and future directions / E. Benetos et al. Journal of Intelligent Information Systems. 2013. Vol. 41, no. 3. P. 407–434. URL: https://link.springer.com/article/10.1007/s10844-013-0258-3.

5. Deep salience representations for f0 estimation in polyphonic music / R. M. Bittner et al. International Society for Music Information Retrieval Conference (ISMIR). 2017. P. 63–70. URL: https://doi.org/10.5281/zenodo.1417937.

6. Böck S., Schedl M. Polyphonic piano note transcription with recurrent neural networks. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2012. P. 121–124. URL: https://doi.org/10.1109/ICASSP.2012.6287832.

7. Residual Shuffle-Exchange Networks for Fast Processing of Long Sequences / A. Draguns et al. Proceedings of the AAAI Conference on Artificial Intelligence. 2021. P. 7245–7253. URL: https://ojs.aaai.org/index.php/AAAI/article/view/16890.

8. Li L. Music Transcription Using Deep Learning. 2017. (Preprint). URL: https://cs229.stanford.edu/proj2017/final-reports/5242716.pdf.

9. Residual Shuffle-Exchange Networks for Fast Processing of Long Sequences / A. Draguns et al. Proceedings of the AAAI Conference on Artificial Intelligence. 2021. P. 7417–7425. URL: https://ojs.aaai.org/index.php/AAAI/article/view/16890.

10. LUMII-Syslab. RSE: Residual Shuffle-Exchange Networks GitHub Repository. URL: https://github.com/LUMII-Syslab/RSE.

11. Convolutional Recurrent Neural Networks for Music Classification / K. Choi et al. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2017. P. 2392–2396. URL: https://dl.acm.org/doi/10.1109/TASLP.2017.2690575.

12. Sigtia S., Benetos E., Dixon S. An End-to-End Neural Network for Polyphonic Piano Music Transcription. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2016. Vol. 24, no. 5. P. 927–939. URL: https://arxiv.org/pdf/1508.01774.

13. Automatic Music Transcription: An Overview / E. Benetos et al. IEEE Signal Processing Magazine. 2019. Vol. 36, no. 1. P. 20–30.

14. Hernandez-Olivan C., al. e. A comparison of deep learning methods for timbre analysis in polyphonic automatic music transcription. Electronics. 2021. Vol. 10, no. 7. P. 810.

15. Sigtia S., Boulanger-Lewandowski N., Dixon S. A Study on LSTM Networks for Polyphonic Music Sequence Modelling. ISMIR 2017. 2017. URL: https://archives.ismir.net/ismir2017/paper/000060.pdf.

16. Kim J., Bello J. Adversarial learning for improved onsets and frames music transcription. Proceedings of the International Society for Music Information Retrieval Conference (ISMIR). 2019. P. 670–677..

17. Hawthorne C., others. Sequence-to-Sequence Piano Transcription with Transformers. Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR). 2021. URL: https://archives.ismir.net/ismir2021/paper/000030.pdf.

18. Team M. Music Transcription with Transformers. URL: https://magenta.tensorflow.org/transcription-with-transformers.

Автори статті

Бай Ярослав – аспірант, Державний університет інформаційно-комунікаційних технологій, Київ, Україна.
ORCID: 0009-0000-0882-4176

Катков Юрій – доктор технічних наук, доцент, Державний університет інформаційно-комунікаційних технологій, Київ, Україна.
ORCID: 0009-0007-1194-4014

Authors of the article

Bai Yaroslav – postgraduate, State University of Information and Communication Technologies, Kyiv, Ukraine.
ORCID: 0009-0000-0882-4176

Katkov Yuriy – Doctor of Sciences (technical), State University of Information and Communication Technologies, Kyiv, Ukraine.
ORCID: 0009-0007-1194-4014