Науковий вісник НЛТУ України
Scientific Bulletin of UNFU
ISSN 1994-7836 (print)
ISSN 2519-2477 (online)
https://doi.org/10.36930/40340514
Article received 21.05.2024
Article accepted 23.05.2024
UDC 004.93
Correspondence author: T. Ya. Ukhachevych taras.y.ukhachevych@lpnu.ua

Т. Я. Ухачевич, Н. О. Кустра
Національний університет "Львівська політехніка", м. Львів, Україна

ДОСЛІДЖЕННЯ ВПЛИВУ ОБРІЗАННЯ ТА ТОНКОГО НАЛАШТУВАННЯ МОДЕЛІ АВТОМАТИЧНОГО РОЗПІЗНАВАННЯ МОВЛЕННЯ НА ЇЇ ТОЧНІСТЬ

Досліджено вплив методів обрізання моделі та тонкого її налаштування на точність автоматичного розпізнавання мовлення ASR (англ. *Automatic Speech Recognition*) для мови з низьким ресурсом. Використану модель "wav2vec2-xls-r-300 m-uk", попередньо навчено на великому багатомовному наборі даних і тонко налаштовано на українському наборі даних із Common Voice. Метод обрізання за L1-нормою було застосовано на різних рівнях (10, 20, 30, 40, 50 %%) без подальшого налаштування, що виявило значне зниження точності (метрика WER (англ. *Word Error Rate*) збільшилася з 18,53 до 35,96 %%, метрика CER (англ. *Character Error Rate*) – з 3,5 до 7,97 %%). Встановлено, що зі збільшенням ступеня обрізання точність моделей поступово знижується, однак подальше тонке налаштування значно покращує продуктивність (метрика WER знизилася до 22,81 %, метрика CER – до 4,55 %). Оцінено вплив кількості епох тонкого налаштування на точність моделі, що показало поступове покращення продуктивності зі збільшенням епох (найкращі результати за 20 епох: метрики WER 22,81 %, CER 4,55 %). З'ясовано, що тонке налаштування здатне частково відновити втрату точності, спричинену обрізанням. Наукова новизна дослідження полягає в комплексному аналізі методів обрізання та тонкого налаштування для низькоресурсних мов на прикладі української мови. Виявлено, що комбіноване використання методів обрізання, перенесення навчання та тонкого налаштування є перспективним для підвищення продуктивності та точності ASR моделей. З'ясовано, що методи обрізання дають можливість зменшити розмір моделі та підвищити її ефективність, що є критичним для пристроїв з обмеженими ресурсами. Охарактеризовано закономірності між ступенем обрізання та ефективністю моделі, що вказують на можливість досягнення балансу між продуктивністю та точністю. З'ясовано, що ефективність моделей для мов з низьким ресурсом значно залежить від кількості та якості навчальних даних, а також від методів їх оброблення. Перспективи подальших досліджень містять аналіз різних методів обрізання, розроблення нових підходів до тонкого налаштування та перенесення навчання з використанням додаткових лінгвістичних ознак. Це сприятиме створенню більш ефективних систем розпізнавання мовлення для мов із низьким ресурсом, що дасть змогу подолати технологічний розрив і покращити доступність мовних технологій у глобальному масштабі.

Ключові слова: нейронні мережі; точність моделі; ефективність; зменшення розміру моделі; перенесення навчання.

Вступ / Introduction

Моделі розпізнавання мовлення є ключовими компонентами сучасних систем, таких як розмовний штучний інтелект, голосові помічники та додатки для транскрипції. Незважаючи на значний прогрес у цій галузі, створення високоточних моделей для мов із низьким ресурсом залишається складним завданням через недостатню кількість навчальних даних. У світовій науковій літературі наведено численні дослідження, спрямовані на оптимізацію моделей розпізнавання мовлення, особливо для мов з великою кількістю ресурсів, таких як англійська та китайська. Проте, мовам з низьким ресурсом приділено значно менше уваги.

Ефективне створення високоточних моделей для мов із низьким ресурсом можливе завдяки використанню методів перенесення навчання, обрізання та тонкого налаштування. Перенесення навчання дає змогу застосовувати вже наявні знання, отримані під час навчання на великих наборах даних для мов з високим ресурсом, та адаптувати модель до мови з низьким ресурсом. Обрізання дає змогу видаляти окремі ваги з моделі нейронної мережі на основі їхньої величини, зменшуючи кількість параметрів без істотного зниження продуктивності. Це зменшує розмір моделі, підвищує її ефективність та знижує енергоспоживання, що є критичним для пристроїв з обмеженими ресурсами, таких як вбудовані системи та пристрої Інтернету речей (IoT). Тонке налаштування, своєю чергою, є процесом додаткового навчання моделі на специфічному наборі даних після обрізання для відновлення або підвищення точності, що дає змогу моделі краще адаптуватися до специфічних мовних або доменних даних, підвищуючи її продуктивність і точність. Дослідження ефектів обрізання, перенесення навчання та тонкого налаштування на моделі розпізнавання мовлення набувають особливої актуальності через необхідність створення ефективних

Інформація про авторів:
Ухачевич Тарас Ярославович, магістр, аспірант, кафедра автоматизованих систем управління.
Email: taras.y.ukhachevych@lpnu.ua; https://orcid.org/0009-0006-5931-4639
Кустра Наталія Омелянівна, канд. техн. наук, доцент, кафедра інформаційних технологій видавничої справи.
Email: nataliia.o.kustra@lpnu.ua; https://orcid.org/0000-0002-3562-2032

Цитування за ДСТУ: Ухачевич Т. Я., Кустра Н. О. Дослідження впливу обрізання та тонкого налаштування моделі автоматичного розпізнавання мовлення на її точність. Науковий вісник НЛТУ України. 2024, т. 34, № 5. С. 104–109.
Citation APA: Ukhachevych, T. Ya., & Kustra, N. O. (2024). Investigating the effect of pruning and fine-tuning of the automatic speech recognition model on its accuracy. *Scientific Bulletin of UNFU*, 34(5), 104–109. https://doi.org/10.36930/40340514

систем для мов з низьким ресурсом. Недостатня кількість анотованих навчальних даних для цих мов призводить до зниження продуктивності моделей, перенавчання та появи помилок, що ускладнює створення ефективних технологій оброблення мови. Аналіз методів оптимізації, таких як обрізання моделей, перенесення навчання та тонке налаштування, стає необхідним для підвищення точності розпізнавання мовлення в умовах обмеженої доступності даних.

Основною ідеєю цього дослідження є аналіз впливу обрізання моделі та подальшого тонкого налаштування на точність систем автоматичного розпізнавання мовлення, розроблених для мов із низьким ресурсом. Використання інтегрованого підходу, який містить методи обрізання, тонкого налаштування та перенесення навчання, забезпечить високу точність та ефективність моделей у реальному часі. В експерименті використано модель, яку було навчено з використанням перенесення навчання на великому наборі даних, а потім тонко налаштовано на малому наборі даних для української мови.

Об'єкт дослідження – системи автоматичного розпізнавання мовлення.

Предмет дослідження – методи обрізання моделі та тонке її налаштування в контексті низькоресурсних мов, що дасть змогу покращити точність та ефективність систем автоматичного розпізнавання мовлення

Мета роботи – дослідити вплив різних ступенів обрізання моделі та подальшого тонкого налаштування на точність автоматичних систем розпізнавання мовлення для мов із низьким ресурсом для підвищення точності, ефективності, зменшення розміру моделі та вимог до обчислювальних ресурсів.

Для досягнення зазначеної мети визначено такі основні *завдання дослідження*:

- проаналізувати ефективність обрізання моделей на різних рівнях, що дасть змогу виявити оптимальні ступені обрізання для збереження максимальної точності;
- оцінити вплив тонкого налаштування на відновлення точності моделей після обрізання, що дасть змогу визначити ефективність та необхідну кількість епох для досягнення найкращих результатів;
- проаналізувати взаємодію методів перенесення навчання з методами обрізання та тонкого навчання для підвищення точності моделей для мов з низьким ресурсом.

Аналіз останніх досліджень та публікацій. Мови з низьким ресурсом створюють унікальні проблеми під час оброблення природної мови NLP (англ. *Neuro-Linguistic Programming*) і автоматичного розпізнавання мовлення ASR (англ. *Automatic Speech Recognition*) через обмежену доступність анотованих даних. Такі методи, як обрізання та тонке налаштування, стали ефективними стратегіями для підвищення продуктивності моделі в цих контекстах. Ці методи дають змогу адаптувати моделі до специфічних мовних або доменних даних, підвищуючи їх точність і надійність.

Дослідники активно шукають способи вирішення проблем, пов'язаних з недостатньою кількістю анотованих наборів даних для мов з низьким ресурсом, що призводить до зниження продуктивності моделей, перенавчання та появи помилок, ускладнюючи створення ефективних технологій оброблення мови для цих мов [11]. Один із підходів полягає у створенні нових наборів даних, що покращують ситуацію з навчанням низькоресурсних мов, як, наприклад, у роботі [8]. Інший підхід передбачає впровадження інноваційних методів машинного навчання, таких як перенесення навчання та тонке налаштування. Використовуючи знання, отримані з мов із високим рівнем ресурсів, перенесення навчання може значно підвищити ефективність у роботі з низько-ресурсними мовами. Тонке налаштування додатково адаптує ці моделі до конкретних завдань або мов, що підтверджується в роботі [12], де вказано важливість цього підходу у сфері ASR.

У роботі [14] досліджено різні методи перенесення навчання для оброблення мовлення та мови, наголошено на їх ефективності в адаптації моделей ASR до нових мов і доменів з обмеженими даними. Висвітлено такі методи, як тонке налаштування моделі та міжмовне перенесення, які мають вирішальне значення для оптимізації ASR для мов із низьким ресурсом.

Однією з ранніх робіт, де запропоновано триетапний метод: навчання щільної мережі, обрізання менш важливих зв'язків та перенавчання мережі для точного налаштування ваг з'єднань, що залишилися, є робота [7]. Дослідження показали, що цей метод не тільки порівнянний зі звичайними моделями глибокого навчання, але й може досягти вищої точності.

У дослідженні [13] розглянуто різні методи обрізання для оптимізації глибоких нейронних мереж шляхом видалення зайвих параметрів. Ці методи є важливими для створення легких моделей ASR, придатних для роботи з мовами з низькими ресурсами, підтримуючи високу продуктивність зі зниженими обчислювальними вимогами.

Інші дослідження також вказують на важливість обрізання моделей. Наприклад, у дослідженні [6] розглянуто вплив обрізання моделей BERT і його наслідки для перенесення її навчання та точного налаштування. У роботі [3] розглянуто неструктуроване обрізання, яке передбачає створення розріджених мереж шляхом ініціалізації вагових коефіцієнтів із розрідженим випадковим розподілом і підтримання розрідженості протягом усього процесу навчання. Результати показують, що розріджені мережі можуть бути більш стійкими до шуму, зберігаючи при цьому конкурентну точність.

Автори дослідження [2] наголошують тим, що нейрони з активними дендритами в мозку можуть надійно та точно розпізнавати візерунки за допомогою невеликої кількості синапсів навіть у шумних умовах. Теоретичні результати та моделювання цього дослідження свідчать про те, що розріджені подання в нейронних мережах дають значні переваги щодо шумостійкості та точності розпізнавання образів. Отримані дані свідчать про те, що використання розрідженості є високоефективною та дієвою стратегією для нейронних мереж.

Отже, хоча більшість перелічених досліджень у сфері оброблення мовлення не розглядали варіанти для мов з низьким ресурсом, результати показують, що комбіноване використання методів обрізання, перенесення навчання та тонкого налаштування є перспективним для підвищення продуктивності та точності моделей. Це вказує на важливість подальших досліджень цих методів у контексті мов з обмеженими ресурсами, де оптимізація моделей є особливо критичною через обмежену доступність анотованих даних. Аналіз цих підходів може сприяти розробленню ефективних систем розпізнавання мовлення для таких мов, подоланню техно-

логічного розриву та покращенню доступності мовних технологій.

Результати дослідження та їх обговорення / Research results and their discussion

Завдання оптимізації. У цьому досліджені поставлено проблему оптимізації моделі автоматичного розпізнавання мови. У нашому контексті критерієм оптимальності є точність розпізнавання мови, яка вимірюється за допомогою метрик WER (англ. *Word Error Rate*) та CER (англ. *Character Error Rate*).

Частоту помилок у слові (WER) обчислюють як відсоток неправильних слів порівняно з загальною кількістю слів у референтному тексті за формулою:

$$WER = \frac{S + D + I}{N}, \quad (1)$$

де: S – кількість замін, D – кількість видалень, I – кількість вставок, а N – загальна кількість слів у референтному тексті. Частота помилок символів (CER) аналогічна WER, але обчислюють на рівні символів за формулою:

$$CER = \frac{S + D + I}{N}, \quad (2)$$

де: S – кількість замін; D – кількість видалень; I – кількість вставок; а N – загальна кількість символів у референтному тексті.

Змінні параметрами є ступенем обрізання моделі (10, 20, 30, 40, 50 %%) та кількість епох під час тонкого налаштування (5, 10, 15, 20 епох).

Основні обмеження містять визначену кількість навчальних даних для української мови (низькоресурсне середовище) та обмеження апаратних ресурсів для тренування та використання моделі.

Математична модель процесу. У цьому дослідженні використано модель з відкритим кодом "wav2vec2-xls-r-300 m-uk", яка базується на фреймворку XLS-R [4]. Цей фреймворк, розроблений на основі wav2vec 2.0 [5], призначений для оброблення великомасштабних багатомовних наборів даних, сприяючи створенню надійних подань мовлення для широкого спектру мов і завдань.

Рисунок ілюструє комплексний підхід до моделі XLS-R, демонструючи потік від немаркованих мовних даних через самоконтрольоване попереднє навчання до точного налаштування для конкретного завдання.

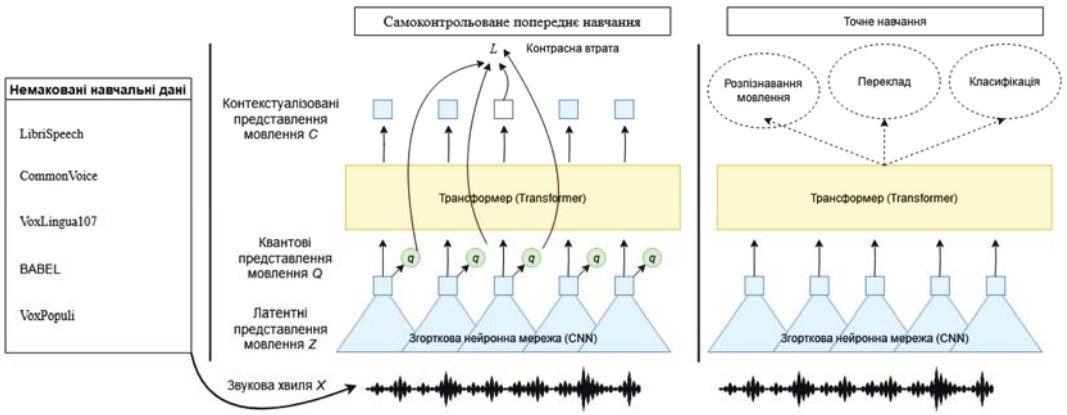

Рисунок. Схема комплексного підходу до моделі XLS-R, демонструючи потік від немаркованих мовних даних через самоконтрольоване попереднє навчання до точного налаштування для конкретного завдання / Figure illustrates the comprehensive approach of the XLS-R model showing the flow from unlabeled speech data through self-supervised pre-training to task-specific fine-tuning

Модель XLS-R містить декілька важливих компонентів, які працюють у тандемі для оброблення необроблених аудіовхідних сигналів у значущі мовні подання. В основі його архітектури знаходиться багатошаровий кодер згорткової нейронної мережі (CNN), який приймає необроблені звукові сигнали ($X = \{x_1, x_2, \ldots, x_n\}$) і перетворює їх на латентні подання мовлення ($Z = \{z_1, z_2, \ldots, z_T\}$). Ці приховані уявлення вловлюють основні акустичні характеристики звукового сигналу.

Ключовим нововведенням у моделі XLS-R є застосування стратегії маскування під час попереднього навчання. Частини прихованих уявлень випадково маскуються, і модель вчиться передбачати ці замасковані сегменти на основі навколишнього контексту. Цей процес покращує здатність моделі вивчати надійні узагальнені характеристики.

Мережа трансформера має важливе значення у контекстуалізації цих прихованих уявлень. Обробляючи послідовність латентних векторів, мережа Трансформер генерує контекстуалізовані подання ($C = \{c_1, c_2, \ldots, c_T\}$), які фіксують залежності по всій аудіопослідовності, уможливлюючи моделі зрозуміти ширший контекст мови.

Для оптимізації процесу навчання в моделі використано контрастну функцію втрат. Основна мета, (L_m), вимагає, щоб модель ідентифікувала справжнє квантоване подання латентного мовлення (q_t) для замаскованого кроку в часі t з набору відволікаючих чинників. Контрастну втрату математично визначаємо як:

$$L_m = -\log\left(\frac{\exp(sim(c_t, q_t)/k)}{\sum_{\tilde{q}\sim Q_t}\exp(sim(c_t, \tilde{q})/k)}\right), \quad (3)$$

де: c_t – це контекстуалізоване подання на кроці часу t, створене мережею Трансформер; q_t – це справжнє квантоване подання прихованої мови для замаскованого кроку часу t, це цільове значення, яке модель повинна правильно передбачити серед $\tilde{q} \in Q_t$ кандитатів квантових подань; k – температурний параметр, який масштабує оцінки подібності. Це допомагає контролювати чіткість розподілу ймовірностей за можливими квантованими поданнями. $sim(a,b)$ – позначає косинусну подібність між двома векторами a і b. У цьому контексті він вимірює подібність між контекстуалізованим поданням c_t і квантованим поданням q_t. Q_t – це набір квантованих подань відволікаючих чинників для замаскованого-

го кроку часу t. Ці відволікаючі чинники є іншими квантованими латентними поданнями мови, які модель повинна відрізняти від справжнього подання q_t.

Окрім цього, модель містить втрату різноманітності кодової книги, щоб заохотити використання всіх її записів:

$$L_d = \frac{1}{GV}\sum_{g=1}^{G} -H(\bar{p}_g) = \frac{1}{GV}\sum_{g=1}^{G}\sum_{v=1}^{V}(\bar{p}_{g,v}\cdot\log\bar{p}_{g,v}), \quad (4)$$

де G – кількість груп у модулі квантування. Кожна група має власний набір записів кодової книги. V – кількість записів кодової книги в кожній групі. Кожен запис представляє окреме квантоване приховане подання. $\bar{p}_g$ – середній розподіл ймовірностей за записами кодової книги для групи g. Це показує, як часто кожен запис кодової книги вибирається протягом усіх часових кроків. $\bar{p}_{g,v}$ – ймовірність вибору v-го запису кодової книги в групі g. $H(\bar{p}_g)$ – ентропія розподілу ймовірностей $\bar{p}_g$. Ентропію обчислюють як:

$$H(\bar{p}_g) = -\sum_{v=1}^{V}\bar{p}_{g,v}\log\bar{p}_{g,v}\ . \quad (5)$$

Функція загальних втрат поєднує ці дві цілі:

$$L = L_m + \alpha L_d\,, \quad (6)$$

де L_m – контрастні втрати, які гарантують, що модель правильно ідентифікує справжні квантовані подання. L_d – втрата різноманітності кодової книги, яка заохочує використання повного діапазону записів кодової книги. α – ваговий коефіцієнт, який врівноважує внески контрастної втрати та втрати різноманітності кодової книги. Цей параметр використано для налаштування відносної важливості втрати різноманітності у функції загальних втрат.

Вхідними даними для моделі XLS-R є необроблена звукова хвиля (waveform), яка складається з послідовності звукових зразків. За допомогою кодувальника функцій і мережі Трансформер модель обробляє ці вхідні дані, щоб створити набір контекстуалізованих подань мовлення. Ці уявлення слугують виходом моделі, фіксуючи відповідні характеристики мовного сигналу, необхідні для подальших завдань.

Процес перенесення навчання. Модель XLS-R піддається процесу перенесення навчання, що містить попереднє навчання на великомасштабних багатомовних наборах даних. Після попереднього навчання модель тонко налаштовується на специфічних завданнях, таких як автоматичне розпізнавання мовлення (ASR), автоматичний переклад мовлення (AST) чи інших завданнях класифікації, містячи ідентифікацію мови та мовця.

Модель "wav2vec2-xls-r-300m-uk", яку використано у цьому дослідженні, власне, і є тонко налаштованою на українському наборі даних Common Voice Corpus 10.0 версією моделі XLS-R.

Методи обрізання та тонкого налаштування. Для оптимізації моделі використовували неструктуроване обрізання за нормою L1 із ступенем обрізання від 10 до 50 %%. Ця техніка дає змогу видаляти вагові коефіцієнти на основі їх абсолютних значень, заохочуючи розрідженість шляхом обнулення ваг із найменшими значеннями. Обрізання L1-норми підтримує продуктивність моделі, зберігаючи менш впливові ваги, і може діяти як метод регуляризації, підвищуючи здатність моделі до узагальнення.

На останньому етапі модель з найбільшим ступенем обрізання (50 %) була піддано тонкому налаштуванню з використанням українського набору даних Common Voice Corpus 10.0, що містить 85 годин українського мовлення. Це робить модель придатною для імітування середовища з низьким ресурсом.

Тренувальні дані. У роботі [4] подано різноманітні джерела немаркованих мовних даних, які використовували для попереднього навчання моделі XLS-R. Джерела містять:

- багатомовний LibriSpeech: 50 тисяч годин прочитаних книг 8 мовами;
- CommonVoice: 7 тисяч годин прочитаної мови 60 мовами;
- VoxLingua107: 6,6 тисячі годин розмови на YouTube 107 мовами;
- BABEL: 1000 годин телефонних розмов 17 мовами;
- VoxPopuli: 372 тисячі годин парламентських виступів 23 мовами.

Дослідження проводилися із точним налаштуванням моделі з використанням українського набору даних Common Voice Corpus 10.0, того самого набору, на якому налаштовувалася модель "wav2vec2-xls-r-300m-uk".

Апаратні і програмні конфігурації. Експерименти проводилися з використанням таких апаратних і програмних конфігурацій:

Апаратне забезпечення:

- центральний процесор: AMD Ryzen 7 5700x;
- графічний адаптер: AMD Radeon 6800XT;
- оперативна пам'ять: 32гб ddr4.

Середовище розроблення: Jupyter notebook.

Програмне забезпечення:

- операційна система: Ubuntu 22.04.4 LTS;
- мова програмування: Python 3.10.12.

Версії бібліотек, які використовували:

- Pytorch 2.3 з ROCm 6.0;
- Transformers 4.40.2;
- Datasets 2.4.0.

Параметри точного налаштування:

- per_device_train_batch_size: 12;
- gradient_accumulation_steps: 6;
- num_train_epochs: 20;
- fp16: True;
- learning_rate: 1 e-4;
- weight_decay: 0,005.

Результати дослідження. Експеримент показує вплив обрізання та тонкого налаштування на модель Wav2vec2. Спочатку необрізана модель показала значення точності WER 18,53 % і CER 3,5 %.

Потім було застосовано неструктуроване обрізання норми L1 до різних ступенів ваг моделі в згорткових і лінійних шарах без точного налаштування та були отримані такі результати (табл. 1):

Табл. 1. Точність моделі за різних рівнів обрізки / Model accuracy at different levels of pruning

Рівень обрізки, %	WER, %	CER, %
10	19,29	3,6
20	19,79	3,85
30	20,52	4,10
40	24,26	4,89
50	35,96	7,97

У міру збільшення обрізки точність моделі поступово знижувалася. Прикметно, у разі обрізання на 50 % точність моделі значно впала до WER 35,96 % і CER

7,97 %. Проте подальше тонке налаштування з використанням українського набору даних із Common Voice Corpus 10.0 покращило точність обрізаної моделі, внаслідок чого WER сягнув 22,81 % і CER 4,55 %. Це демонструє, що в той час як обрізання знижує точність, точне налаштування може частково відновити продуктивність навіть за умов низькоресурсної мови (табл. 2).

Табл. 2. Точність моделі після тонкого налаштування з різною кількістю епох тренування / Model accuracy after fine-tuning with different number of training epochs

Кількість епох тренування	WER, %	CER, %
5	25,48	5,16
10	24,00	4,81
15	23,24	4,66
20	22,81	4,55

Результати дослідження вказують на важливі особливості ефективності методів обрізання та тонкого налаштування моделей мовлення в умовах низького ресурсу. По-перше, значне збільшення WER і CER після застосування 50 % обрізання неструктурованої норми L1 вказує на те, що агресивне обрізання може істотно погіршити продуктивність моделі, особливо в задачах розпізнавання мовлення, де акустичні деталі мають вирішальне значення.

Проте подальше доопрацювання обрізаної моделі може пом'якшити деякі з цих несприятливих ефектів. Тонке налаштування з використанням невеликого українського набору даних привело до покращень як у WER, так і в CER, хоча модель не відновила початкові рівні продуктивності. Це часткове відновлення вказує потенціал тонкого налаштування як важливого етапу для адаптації обрізаних моделей для мов із низьким ресурсом, допомагаючи відновити втрачену точність.

Окрім цього, результати показують вплив різної кількості епох навчання під час процесу тонкого налаштування. Зі збільшенням кількості епох продуктивність обрізаної моделі поступово покращувалася, що свідчить про можливість подальшого підвищення точності моделі через розширення процесу тонкого налаштування.

Загалом, результати свідчать про те, що хоча обрізання може негативно вплинути на продуктивність моделі, поєднання обрізання з тонким налаштуванням може забезпечити більш збалансований компроміс між ефективністю та точністю. Цей підхід є особливо корисним у середовищах з обмеженими ресурсами, де доступність великих навчальних даних обмежена, але існує потреба в ефективних моделях мовлення.

Обговорення результатів дослідження. У роботі [1] оцінено вплив обрізання моделі на продуктивність машинного перекладу для мови з низьким ресурсом. Результати показують, що обрізання зменшує розмір моделі та обчислювальні вимоги без істотної шкоди для продуктивності під час оброблення частих речень. Однак агресивне обрізання може погіршити результати, тому важливо збалансувати рівень розрідженості для уникнення негативних наслідків. Це вказує на потенціал тонкого налаштування в пом'якшенні негативних ефектів обрізання та оптимізації моделей для низькоресурсних мовних додатків.

У роботі [15] досліджено вплив обрізання на багатомовну модель ASR. Результати показують, що навіть складні алгоритми обрізання негативно впливають на точність моделі. Це свідчить про необхідність застосування інноваційних методів скорочення та оптимізації у комплексі для розроблення ефективних моделей ASR у середовищах з обмеженими ресурсами.

У роботі [10] розглянуто різні методи скорочення, вказуючи на важливість як статичних, так і динамічних методів для підтримки продуктивності моделі у разі зниження обчислювальної складності. Авторам вдалося показати, що агресивність обрізання та точність моделі є компромісом, а тонке налаштування або додаткові етапи навчання можуть відновити продуктивність. Вони також акцентують увагу на квантуванні як додатковому методі, що ілюструє широкий ландшафт стратегій обрізання моделей для ефективного розгортання нейронних мереж.

У роботі [9] підтверджено факт зниження продуктивності моделі після обрізання та досліджено стратегії точного налаштування. Результати показують, що достатньо інтенсивне точне налаштування може не тільки відновити, але й покращити точність моделі.

У роботі [16] досліджено методи переносу навчання та тонкого налаштування для підвищення продуктивності ASR у мовах з низьким ресурсом. Це співвідноситься з нашим підходом, який використовує невеликий український набір даних для точного налаштування обрізаних моделей, демонструючи покращення показників точності.

Отже, за результатами виконаної роботи можна сформулювати такі наукову новизну та практичну значущість результатів дослідження.

Наукова новизна отриманих результатів дослідження – розроблено підхід до поєднання неструктурованого обрізання та тонкого налаштування для моделі Wav2vec2 у контексті низькоресурсних мов, який показав, що агресивне обрізання може значно знизити точність моделі, але подальше тонке налаштування з використанням невеликого набору даних відновлює продуктивність моделі, вказуючи на важливість збалансованого застосування цих методів для оптимізації моделей ASR.

Практична значущість результатів дослідження – можливості застосування отриманих результатів для розроблення ефективних методів обрізання та тонкого налаштування моделей розпізнавання мовлення в умовах з обмеженою кількістю навчальних даних, що дасть змогу покращити продуктивність таких моделей, знизити вимоги до обчислювальних ресурсів та підвищити точність розпізнавання мовлення, особливо в середовищах з обмеженими навчальними даними.

Висновки / Conclusions

Досліджено методи аналізу впливу різних ступенів обрізання та тонкого налаштування на продуктивність автоматичних систем розпізнавання мовлення для мов із низьким ресурсом, що дало змогу покращити точність та ефективність систем автоматичного розпізнавання мовлення, зменшити розмір моделі та її вимоги до обчислювальних ресурсів. За результатами проведеного дослідження можна зробити такі основні висновки.

1. Досліджено вплив неструктурованого обрізання на модель Wav2vec2, встановлено, що воно значно знижує кількість параметрів моделі, але водночас знижує точність, що проявляється у збільшенні частоти помилок у словах (WER) і частоти помилок символів (CER).

2. Встановлено, що точне налаштування обрізаних моделей з великою кількістю епох навчання демонструє потенціал для відновлення продуктивності.
3. Виявлено, що більша кількість епох тонкого налаштування підвищує точність моделі, вказуючи на важливість цього процесу для адаптації обрізаних моделей для мов із низьким ресурсом.
4. Подано стратегічне поєднання обрізання та точного налаштування як збалансований компроміс між ефективністю та точністю моделі, що є вигідним для середовищ із низьким ресурсом.

Перспективи подальших досліджень містять:

- дослідження різних методів обрізання, зокрема структурованого обрізання, для подальшої оптимізації моделей;
- розроблення нових підходів до тонкого налаштування, що дадуть змогу ще краще відновлювати продуктивність обрізаних моделей;
- аналіз можливостей перенесення навчання з використанням додаткових лінгвістичних ознак для підвищення точності моделей для мов із низьким ресурсом;
- аналіз впливу різних обсягів та якості навчальних даних на ефективність обрізання та тонкого налаштування, з метою виявлення оптимальних конфігурацій для різних мовних контекстів.

Ці напрями досліджень можуть сприяти подальшому розвитку ефективних та точних систем розпізнавання мовлення для мов із низьким ресурсом, підвищуючи їх доступність і корисність у глобальному масштабі.

References

1. Ahia, O., Kreutzer, J., & Hooker, S. (2021). The Low-Resource Double Bind: An Empirical Study of pruning for Low-Resource Machine Translation. *ArXiv* (*Cornell University*). https://doi.org/10.48550/arxiv.2110.03036
2. Ahmad, S., & Hawkins, J. (2016). How do neurons operate on sparse distributed representations? A mathematical theory of sparsity, neurons and active dendrites. *ArXiv* (*Cornell University*). https://doi.org/10.48550/arxiv.1601.00720
3. Ahmad, S., & Scheinkman, L. (2019). How can we be so dense? The benefits of using highly sparse representations. *ArXiv* (*Cornell University*). https://doi.org/10.48550/arxiv.1903.11257
4. Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., Patrick, V. P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., & Auli, M. (2021). XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. *ArXiv* (*Cornell University*). https://doi.org/10.48550/arxiv.2111.09296
5. Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *ArXiv* (*Cornell University*). https://doi.org/10.48550/arxiv.2006.11477
6. Gordon, M. A., Duh, K., & Andrews, N. (2020). Compressing BERT: Studying the Effects of weight pruning on Transfer Learning. *ArXiv* (*Cornell University*). https://doi.org/10.48550/arxiv.2002.08307
7. Han, S., Pool, J., Tran, J., & Dally, W. J. (2015). Learning both Weights and Connections for Efficient Neural Networks. *ArXiv* (*Cornell University*). https://doi.org/10.48550/arxiv.1506.02626
8. Kimanuka, U., Maina, C. wa, & Büyük, O. (2024). Speech recognition datasets for low-resource Congolese languages. *Data in Brief*, 52. https://doi.org/10.1016/j.dib.2023.109796
9. Le, D. H., & Hua, B. (2021). Network pruning that matters: A case study on retraining variants. *ArXiv* (*Cornell University*). https://doi.org/10.48550/arxiv.2105.03193
10. Liang, T., Glossner, J., Wang, L., Shi, S., & Zhang, X. (2021 b). Pruning and Quantization for Deep Neural Network Acceleration: A survey. *ArXiv* (*Cornell University*). https://doi.org/10.48550/arxiv.2101.09671
11. Mirela. (2024). Low-resource languages: A localization challenge. POEditor Blog. URL: https://poeditor.com/blog/low-resource-languages/
12. Rista, A., & Kadriu, A. (2020). Automatic Speech Recognition: A Comprehensive Survey. *SEEU Review*, 15(2), 86–112. https://doi.org/10.2478/seeur-2020-0019
13. Vadera, S., & Ameen, S. (2022). Methods for Pruning Deep Neural Networks. Access, 10, 63280–63300. https://doi.org/10.1109/access.2022.3182659
14. Wang, D., & Thomas Fang Zheng. (2015). Transfer Learning for Speech and Language Processing. *ArXiv* (*Cornell University*). https://doi.org/10.48550/arxiv.1511.06066
15. Yang, M., Tjandra, A., Liu, C., Zhang, D., Le, D., Hansen, J. H. L., & Kalinli, O. (2022). Learning ASR pathways: A sparse multilingual ASR model. *ArXiv* (*Cornell University*). https://doi.org/10.48550/arxiv.2209.05735
16. Yu, C., Chen, Y., Li, Y., Kang, M., Xu, S., & Liu, X. (2019). Cross-Language End-to-End Speech Recognition Research based on transfer Learning for the Low-Resource Tujia Language. *Symmetry*, *11*(2), 179. https://doi.org/10.3390/sym11020179

T. Ya. Ukhachevych, N. O. Kustra

Lviv Polytechnic National University, Lviv, Ukraine

INVESTIGATING THE EFFECT OF PRUNING AND FINE-TUNING OF THE AUTOMATIC SPEECH RECOGNITION MODEL ON ITS ACCURACY

Automatic speech recognition (ASR) models are essential in modern technologies such as conversational AI, voice assistants, and transcription applications, yet achieving high accuracy for low-resource languages remains challenging due to insufficient training data. This study explores the effects of model pruning and fine-tuning on the performance of ASR systems for low-resource languages, focusing on Ukrainian. We employed an integrated approach combining transfer learning, pruning, and fine-tuning, using the "wav2vec2-xls-r-300m-uk" model that was initially pre-trained on large-scale multilingual datasets and then fine-tuned on a smaller Ukrainian dataset. Pruning was applied at various levels (10, 20, 30, 40, and 50 %%) without prior fine-tuning. Our results show that as pruning levels increased, model accuracy decreased significantly, with the highest pruning level (50 %) resulting in a WER of 35.96 % and a CER of 7.97 %. However, subsequent fine-tuning using Ukrainian data improved the pruned model accuracy, achieving a WER of 22.81 % and a CER of 4.55 % after 20 epochs. This indicates that while aggressive pruning negatively impacts performance, fine-tuning can substantially recover accuracy. Consequently, this approach demonstrates a balanced trade-off between model efficiency and accuracy, crucial for resource-constrained environments. The benefits of this study are evident in its potential to enhance ASR systems for low-resource languages, providing a framework for optimizing model performance with limited data. Moreover, the findings underscore the importance of fine-tuning as a crucial step in adapting pruned models to specific low-resource contexts. Future research should include exploring different pruning methods, developing new fine-tuning approaches, and employing transfer learning with additional linguistic features. These advancements will contribute to the development of more efficient ASR systems for low-resource languages, helping to bridge the technological gap and improve the accessibility of language technologies globally.

Keywords: Neural networks; model accuracy; efficiency; size reduction; transfer learning.