УДК 004.52

МОДЕЛЮВАННЯ СИСТЕМИ АВТОМАТИЧНОГО РОЗПІЗНАВАННЯ УКРАЇНСЬКОГО МОВЛЕННЯ ЗА ДОПОМОГОЮ ТРАНСФОРМЕРА

В.С. Тарасенко, магістрант
Київський національний університет технологій та дизайну
М.І. Гольдберг, кандидат технічних наук, доцент
Київський національний університет технологій та дизайну

Ключові слова: автоматичне розпізнавання мовлення, глибинне навчання, трансформер, самоувага, перенесення навчання, Whisper.

Спектр використання автоматичного розпізнавання мовлення дуже широкий: створення нотаток із бізнес зустрічей у корпоративному світі, автоматична генерація субтитрів для контенту в індустрії розваг, інтеграція у пристрої для допомоги людям з обмеженими можливостями тощо. Обробка природної мови полягає в пошуку зв'язків між складовими частинами мови. Попередня обробка даних та вилучення ознак є важливими етапами створення будь-якої моделі.

Традиційні підходи створення моделей будувалися на основі прихованих моделей Маркова та сумішей Гауса. Останніми роками традиційні підходи все частіше включали різні техніки глибинного навчання, а на даний момент, найбільш точні моделі відмовилися від традиційних підходів і використовують тільки техніки глибинного навчання, такі як рекурентність, самоувага, декодер-кодер.

У роботі розглядаються сучасні та класичні архітектури систем автоматичного розпізнавання мовлення. Особлива увага приділяється архітектурі трансформер, адже вона досягає найкращої точності та стійкості. Трансформер побудований на підході Seq2Seq: складається з кодера та декодера. Декодер фіксує контекст вхідної послідовності та передає його кодеру, що генерує вихідну послідовність на основі контексту. Трансформер уникає рекурентності, а для того, щоб запам'ятовувати зв'язки між словами вхідного речення використовує механізм самоуваги. Самоувага дозволяє обробляти усі слова вхідної послідовності одночасно, а не один за одним, як під час рекурентності. Це допомагає пришвидшити процес прогнозування і уникнути проблеми зникнення/вибуху градієнта. Позиційні кодування потрібні для фіксації порядку вхідних даних.

Перший крок створення моделі – збір набору даних для тренування. Поєднання декількох різних наборів даних допоможе моделі уникнути перенавчення (overfitting), адже дані належать різним джерелам і мовникам. За основу для нашої моделі ми візьмемо попередньо навчену модель та скористаємося технікою перенесення навчання. Мотивацію

перенесення навчання можна знайти в ідеї «Вчимося вчитися» (NIPS 95), що стверджує, що навчання tabula rasa часто обмежене. Природні мови для задачі автоматичного розпізнавання мовлення розрізняють на дві категорії: мови з високим ресурсом та мови з низьким ресурсом (за кількістю даних для тренування в свободному доступі). Англійська є прикладом мови з високим ресурсом, українська – з низьким. В нашому випадку перенесення навчання буде відбуватися з англійської мови на українську.

Одним з найкращих представників архітектури трансформер у галузі автоматичного розпізнавання мовлення є модель компанії OpenAI під назвою Whisper (WSPSR – Web-scale Supervised Pre-training for Speech Recognition). Ця модель є сильним кандидатом для роботи з мовами з низьким ресурсом, як-от українська. Модель підтримує частоту дискретизації аудіодоріжки 16kHz, дуже важливо щоб вхідні дані дотримувалися цієї вимоги. Whisper тренований на аудіофайлах довжиною до 30 секунд і не може приймати більш довгі дані. Для нашої моделі була вибрана конфігурація Whisper під назвою small (конфігурації відрізняються за розміром: кількістю тренувальних параметрів, кількістю шарів, шириною шарів, кількістю голов механізму самоуваги).

Однією з найбільших проблем під час цього процесу було керування величезним обсягом даних. У підсумку модель було протреновано на 100 малих піднаборах даних (кожний набор містить 7,7 тисяч тренувальних та 976 валідаційних екземплярів) по одному епоху (повного проходження піднабору, одна партія за раз) на кожний піднабір.

Після тренування наша модель була порівняна з базовим Whisper. Показник WER був знижений на 70-90% після тренування. Потім ми порівняли точність нашої моделі на тестових даних з результатами існуючої моделі VOSK, представника класичної архітектури – прихованих моделях Маркова. Наша модель перевершила модель VOSK на майже усіх ділянках. Середній WER нашої моделі на 15% нижчий за VOSK.

Для демонстрації роботи моделі був створений веб-додаток.

Список використаних джерел

1. Ashish Vaswani, et al. Attention is all you need.: Advances in Neural Information Processing Systems, vol. 30, 2017.

2. Wang, Dong, et al. Transfer Learning for Speech and Language Processing.: In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, 2–5. IEEE.

3. Josue Batista. Learn OpenAI Whisper: Transform your understanding of GenAI through robust and accurate speech processing solutions.: Packt Publishing., 2024, 372p.