УДК 004.8

Є.М Лучко
О.С. Городецька

СИСТЕМА ГОЛОСОВОГО ВВЕДЕННЯ ТЕКСТУ З ВИКОРИСТАННЯМ НЕЙРОМЕРЕЖІ

Вінницький національний технічний університет

Анотація
У даній роботі розглядається доцільність та переваги застосування нейромережі з трансформерною архітектурою в задачачах розпізнавання природної мови в рамках реалізації системи голосового введення тексту.
Ключові слова: голосове введення тексту, нейромережі, трансформерна архітектура

Abstract
This work considers the feasibility and advantages of using a neural network with a transformer architecture in natural speech recognition tasks as part of the implementation of a speech input system.
Keywords: speech input, neural networks, transformer architecture

Вступ

Галузь обробки та розпізнавання природної мови зазнає стрімкого розвитку протягом останнього десятиліття, що дозволяє доповнити досвід взаємодії людини з комп'ютерними пристроями. Одна з технологій, що має важливе значення в цьому контексті, є голосове введення тексту.

Голосове введення тексту представляє собою процес перетворення вимовлених людиною слів на текст за допомогою спеціальних алгоритмів та програмного забезпечення. Ця технологія знаходить широке застосування в різноманітних сферах, включаючи інформаційні технології, медицину, освіту та бізнес.

Одним з найважливіших факторів, що призвели до зростання інтересу до голосового введення тексту, є значний прогрес у сфері штучного інтелекту, особливо в глибокому навчанні та нейронних мережах. Ці технології дозволяють створювати дедалі точніші та ефективніші системи розпізнавання мови, що робить голосове введення тексту більш доступним і зручним для користувачів.

Обґрунтування вибору нейромережі

Найбільш сучасною та ефективною архітектурою для розпізнавання природної мови є трансформер. Завдяки здібності до паралельної обробки даних, ця архітектура дозволяє значно швидше проводити обчислення, на відміну від рекурентних нейронних мереж, які обробляють дані послідовно. Також визначною характеристикою архітектури є механізм уваги, що дозволяє знаходити найбільш важливу інформацію в різних частинах вхідних даних та розуміти її поточний контекст. Такі параметри архітектури дозволяють ефективно виконувати розпізнавання голосу в реальному часі, що дозволяє використання подібних моделей для голосового введення тексту. [1]

Однією з значущих перешкод в застосуванні нейромереж для цілей розпізнавання мови була складність створення якісного набору даних, що робило тренування мережі значно повільніше та дорожче. Особливо нагальною проблема була під час роботи з мовами, що мають малу кількість носіїв та відповідно малу кількість дослідників та оброблених ними даних, що придатні для навчання нейромереж. Для вирішення цієї проблеми використовується метод неконтрольованого міжмовного представлення (uncontrolled cross-lingual representation, UCR). Унікальність UCR полягає в здатності знаходити збіжності між мовами навіть без навчальних пар, що дозволяє переносити знання з однієї мови на іншу та використовувати мінімально оброблені набори даних. Ці представлення можуть бути використані для вирішення різноманітних задач обробки природної мови на різних мовах, що особливо корисно для сфер діяльності де складно анотувати дані. [2]

Прикладом такої нейромережі є Whisper, багатомовна мережа на архітектурі трансформер, яка є

основним елементом в системі голосового введення. На даний момент ця модель є найточнішою та підтримує роботу з українською мовою. Вхідні дані не потребують попередньої обробки, адже ця модель була розроблена з ідеєю роботи «з коробки», тобто без попередньої підготовки вхідних даних, зміни гіперпараметрів, донавчання моделі на додаткових даних. [3]

Система введення працює таким чином: спочатку стартуємо потокову передачу аудіодаих (використовуючи pyaudio), після чого потокові дані передаються на вхід нейромережі, на виході отримуємо готовий текст, що вводиться в активне текстове поле.

Висновки

Враховуючи значні переваги нейромереж з архітектурою трансформер при роботі з послідовними даними, а конкретно при розпізнаванні людської мови, можна зробити висновок про доцільність їх застосування в системі голосового введення.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Attention Is All You Need [Електронний ресурс]. – Режим доступу: https://arxiv.org/abs/1706.03762
2. Robust Speech Recognition via Large-Scale Weak Supervision [Електронний ресурс]. – Режим доступу: https://arxiv.org/abs/2212.04356
3. Unsupervised Cross-lingual Representation Learning at Scale [Електронний ресурс]. – Режим доступу: https://arxiv.org/abs/1911.02116

Лучко Євген Миколайович — студент групи 1КІ-20б, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, e-mail: eweismann03@gmail.com

Городецька Оксана Степанівна — кандидат технічних наук, доцент кафедри обчислювальної техніки, Вінницький національний технічний університет, Вінниця, e-mail: gorodeczka.o.s@vntu.edu.ua

Luchko Yevhen M. — student of the 1KI-20b group, faculty of Information Technologies and Computer Engineering, Vinnytsia National Technical University, Vinnytsia e-mail: eweismann03@gmail.com

Horodetska Oksana S. — Candidate of Technical Sciences, Associate Professor of the Department of Computer Engineering, Vinnytsia National Technical University, Vinnytsia, e-mail: gorodeczka.o.s@vntu.edu.ua