УДК 004.934

ИСПОЛЬЗОВАНИЕ ТРАНСФОРМЕРОВ В СИСТЕМАХ РАСПОЗНАВАНИЯ РЕЧИ

Мурадов М.Т., Заведующий кафедрой информационных систем Института Телекоммуникаций и информатики Туркменистан , г. Ашхабат.
Научный руководитель: Оразгельдиев А.Х, к.т.н.,
Института Телекоммуникаций и информатики Туркменистан , г. Ашхабат.

Аннотация. Данная работа посвящена исследованию применения трансформеров в системах автоматического распознавания речи (ASR). Рассматриваются основные преимущества использования трансформеров, такие как способность эффективно обрабатывать длинные последовательности данных, учитывать контекст и обеспечивать высокую точность в условиях многозначности и шума. Особое внимание уделено популярным моделям, таким как Wav2Vec 2.0 и Conformer, которые используют механизмы самообучения и внимания для улучшения качества распознавания речи. Также обсуждаются проблемы, связанные с вычислительными затратами и потребностью в большом объеме аннотированных данных, а также возможности дальнейшего развития трансформеров в этой области. Работа подчеркивает важность трансформеров для создания более точных и устойчивых ASR-систем, способных эффективно работать в реальных условиях.

Ключевые слова: Трансформеры, Распознавание речи (ASR), Механизм внимания (Attention), Модели Wav2Vec 2.0, Модели Conformer, Самообучение (Self-supervised learning), Контекстная зависимость, Акустические признаки, Шумоустойчивость, Многозначность, Точность распознавания, Нейронные сети, Обработка аудио, Алгоритмы машинного обучения, Речевые технологии.

Использование трансформеров в системах автоматического распознавания речи (ASR, Automatic Speech Recognition) стало важным шагом в развитии этой области. Ранее в таких системах доминировали скрытые марковские модели (HMM), а также рекуррентные нейронные сети (RNN), включая LSTM и GRU. Однако трансформеры, благодаря их способности эффективно обрабатывать длинные последовательности данных и учитывать контекст, существенно улучшили точность и гибкость распознавания речи. Давайте более подробно рассмотрим, как именно трансформеры применяются в ASR.

1. Основы трансформеров в ASR.

Трансформеры — это тип нейронных сетей, которые были впервые представлены в статье "Attention is All You Need" (2017) от Vaswani и соавторов. Их основное преимущество заключается в механизме self-attention, который позволяет моделям учитывать все элементы последовательности одновременно (в отличие от RNN, которые обрабатывают данные последовательно). Это особенно важно для распознавания речи, так как

аудиопоследовательности часто содержат сложные и длинные временные зависимости.

Основные преимущества трансформеров в ASR:

Self-attention (внимание к себе): позволяет модели эффективно учитывать взаимосвязи между всеми элементами последовательности, независимо от их позиции. Это особенно важно для учета контекста в речевых данных.

Параллельная обработка: В отличие от RNN, которые требуют последовательной обработки данных (что делает обучение и инференс более медленным), трансформеры могут обрабатывать данные параллельно, что значительно ускоряет обучение.

В основе любой речевой технологии лежит так называемый «engine» или ядро программы - набор данных и правил, по которым осуществляется обработка данных. В зависимости от назначения этого ядра различают TTS и ASR engine. TTS (Text-to-Speech) engine предоставляет возможность синтеза речи по тексту, а ASR (Automatic Speech Recognition) engine - для распознавания речи. Существует несколько крупных производителей, занимающихся созданием ASR ядер и среди них такие компании, как SPIRIT, Advanced Recognition Technologies, IBM. (2)

Контекстуальная зависимость: Трансформеры могут учитывать, как локальные, так и глобальные зависимости в аудио последовательности. Это помогает распознавать слова и фразы, которые могут быть нечетко произнесены.

2. Роль трансформеров в распознавании речи.

Трансформеры применяются на разных этапах процесса распознавания речи. Рассмотрим подробно, на практике.

2.1 Преобразование аудио сигнала в текстовой формат.

Речь представляет собой последовательность звуковых волн, которая должна быть преобразована в текст. Это преобразование можно разбить на несколько этапов:

Извлечение признаков из аудио: Аудио сигнал сначала проходит через систему извлечения признаков, которая преобразует его в более компактное и информативное представление, например, в мел-спектрограмму (Mel-spectrogram). Эти признаки отражают интенсивность звука в различных частотных диапазонах.

Обработка с использованием трансформеров: После того как аудио сигнал преобразован в признаки, эти данные передаются в модель на основе трансформеров, которая анализирует всю последовательность признаков с учётом контекста и на основе само внимания (self-attention) определяет, какие части аудио соответствуют каким словам или фразам.

В отличие от классических методов, которые обрабатывают данные последовательно (например, используя RNN или LSTM), трансформеры обрабатывают всю последовательность сразу, что позволяет быстрее учитывать глобальные зависимости и решать проблемы, связанные с длинными временными зависимостями.

Транскрипция и предсказание: На основе этой информации модель генерирует текст, который соответствует в аудио звукам. Для повышения точности, особенно в сложных языках или условиях, такие модели используют предсказание на основе контекста (например, BERT или GPT) для завершения фраз и корректировки возможных ошибок в распознавании.

Сбор данных является основой всех процессов работы с Big Data. Здесь важно рассказать о различных источниках данных, которые могут быть использованы, и о методах их сбора. Важно упомянуть о сложностях и проблемах, которые могут возникнуть при сборе данных, таких как неполнота, шумность данных и проблемы с качеством.

2.2 Примеры использования трансформеров в ASR.

Некоторые популярные модели, использующие трансформеры для распознавания речи, включают:

Wav2Vec 2.0 (Facebook AI): Это одна из самых известных моделей для ASR, которая использует трансформеры. Модель Wav2Vec 2.0 обучается на необработанных аудиофайлах, а затем до обучается на данных с транскрипциями. Основная идея заключается в том, чтобы извлечь скрытые признаки из необработанных аудио данных с минимальными метками, что значительно уменьшает потребность в больших аннотированных данных для обучения.

Conformer (Google): Conformer — это модель, которая сочетает в себе преимущества трансформеров и сверточных нейронных сетей (CNN). CNN используются для извлечения локальных признаков из аудио, а трансформеры обрабатывают длинные последовательности и контекст на уровне слов. Это позволяет достигать высокой точности и устойчивости к шуму.

DeepSpeech (Mozilla): DeepSpeech не является чисто трансформерной моделью, она использует идеи, схожие с теми, что применяются в трансформерах, для обработки аудио данных и их преобразования в текст. Это включает использование многослойных нейронных сетей и механизмов внимания для улучшения качества распознавания.

3. Как трансформеры улучшают точность ASR.

Улучшенная способность учитывать контекст: Трансформеры превосходно справляются с задачами, где важно учитывать контекст. Это особенно важно при распознавании длинных фраз или предложений, когда значение слов зависит от всего контекста.

Снижение ошибки при распознавании схожих звуков: Трансформеры могут лучше различать схожие по звучанию слова, так как они анализируют весь контекст, а не только текущие звуки.

Устойчивость к шуму: Трансформеры могут быть обучены работать в условиях шума и искажений, что делает их идеальными для распознавания речи в реальных условиях, например, в шумных помещениях или при разговоре с акцентами.

Обработка многозначности: Модели на базе трансформеров помогают разрешать многозначность слов и фраз, что является важной задачей для точного распознавания речи.

4. Проблемы и вызовы.

Хотя трансформеры существенно улучшили качество распознавания речи, есть и несколько проблем:

Высокие вычислительные ресурсы: Трансформеры требуют значительных вычислительных мощностей для обучения. Это особенно важно, когда речь идет о мобильных устройствах или системах с ограниченными ресурсами.

Нужда в большом объеме аннотированных данных: Хотя пред обучение моделей на больших объемах данных без транскрипций помогает снизить зависимость от аннотированных данных, для достижения высокой точности все равно требуется большое количество меток.

Сложности с адаптацией к новым языкам или акцентам: Несмотря на свою гибкость, модели на основе трансформеров могут иметь сложности с адаптацией к новым языкам или акцентам, особенно если они не были обучены на соответствующих данных.

Использование трансформеров в системах распознавания речи открыло новые горизонты в точности и гибкости распознавания. Механизмы внимания и способность обрабатывать долгосрочные зависимости позволяют этим моделям превосходно справляться с сложными и многозначными речевыми данными, улучшая качество автоматического преобразования аудио в текст. Однако, несмотря на впечатляющие успехи, существует ряд вызовов, включая необходимость в вычислительных ресурсах и большом объеме аннотированных данных для эффективного обучения.

Список литературы:

1. ОД Чарыева, НЧ Реджепмырадова, МТ Мурадов Модели и алгоритмы в системах анализа речевых сигналов Издательство «Всемирный ученый» 2023 156-163
2. Сона Назаровна Назарова, Максат Мырадов Современные технологии распознавания речи Научное издательство «Наука и мировоззрение» 2024 301-305 с
3. Ирина Курбангельдыевна Овездурдыева, Максат Мырадов Технологии работы с большими данными “big data”: сбор, хранение и обработка больших данных Наука и мировоззрение» 2024 70-74 с

USING TRANSFORMERS IN SPEECH RECOGNITION SYSTEMS

Muradov M.T., Head of the Department of Information Systems, Institute of Telecommunications and Informatics of Turkmenistan, Ashgabat.
Scientific supervisor: Orazgeldyev A.Kh., C.T.S
Institute of Telecommunications and Informatics of Turkmenistan, Ashgabat.

Abstract. This work is dedicated to the study of the application of transformers in automatic speech recognition (ASR) systems. The main advantages of using transformers are discussed, such as their ability to efficiently process long data sequences, consider context, and ensure high accuracy in conditions of ambiguity and noise. Special attention is given to popular models like Wav2Vec 2.0 and Conformer, which utilize self-supervised learning and attention mechanisms to improve speech recognition quality. The paper also addresses challenges related to computational costs and the need for large amounts of annotated data, as well as the potential for further development of transformers in this field. The work emphasizes the importance of transformers for creating more accurate and robust ASR systems capable of operating effectively in real-world conditions.

Keywords: Transformers, Speech Recognition (ASR), Attention Mechanism, Wav2Vec 2.0 Models, Conformer Models, Self-supervised Learning, Context Dependency, Acoustic Features, Noise Robustness, Ambiguity, Recognition Accuracy, Neural Networks, Audio Processing, Machine Learning Algorithms, Speech Technologies.