УДК 004.522

АРХИТЕКТУРА TRANSFORMER ДЛЯ ПРЕОБРАЗОВАНИЯ ГОЛОСА В ТЕКСТ

М. Т. Мырадов
Заведующий кафедры
«Информационные системы»
Институт Телекоммуникаций и
информатики Туркменистана
maksat.myradow.92@mail.ru

М.Т. Мырадов
Окончил Институт телекоммуникаций и информатики Туркменистана. Область научных интересов распознавания речи, искусственный интеллект, защита данных

Аннотация: Архитектура Transformer успешно применяется для преобразования голоса в текст благодаря своей способности эффективно обрабатывать последовательности данных. Модель использует механизмы внимания (self-attention) для выявления зависимости между элементами входных данных, что повышает точность распознавания речи. Такой подход позволяет системе обучаться на больших объемах аудиоданных и добиваться высокой производительности в реальных приложениях.

Ключевые слова: Архитектура Transformer, преобразование речи в текст, внимание, распознавание речи.

Введение. Архитектура Transformer была впервые представлена компанией Google в 2017 году в статье под названием «Attention Is All You Need». Эта модель эффективно используется в различных областях языка, включая системы автоматического распознавания речи (ASR). **“Attention Is All You Need”.** Доминирующие модели последовательной трансдукции основаны на сложных рекуррентных или сверточных нейронных сетях в конфигурации кодер-декодер. Наиболее эффективные модели также связывают кодер и декодер через механизм внимания. Мы предлагаем новую простую сетевую архитектуру, Transformer, основанную исключительно на механизмах внимания, полностью обходясь без рекуррентности и сверток. Эксперименты с двумя задачами машинного перевода показывают, что эти модели превосходят по качеству, будучи более параллелизуемыми и требуя значительно меньше времени на обучение. Наша модель достигает 28,4 BLEU в задаче перевода с английского на немецкий язык WMT 2014, улучшая существующие лучшие результаты, включая ансамбли, более чем на 2 BLEU. В задаче перевода с английского на французский язык WMT 2014 наша модель устанавливает новый современный показатель BLEU для одной модели в 41,8 после обучения в течение 3,5 дней на восьми графических процессорах, что составляет небольшую часть затрат на обучение лучших моделей из литературы. Мы показываем, что Transformer хорошо обобщается и на другие задачи, успешно применяя его для анализа английской электоральной аудитории как с большими, так и с ограниченными обучающими данными**.**

Основы модели трансформатора

Модель представляет собой высокоскоростную и универсальную модель, разработанную на основе архитектуры Transformer. Трансформатор состоит из кодера и декодера. Кодер обрабатывает входные данные (мелодию голоса, дикцию и структуру слова), а декодер используется для интерпретации этих данных.

Модель:

1 Кодер: спектрально кодирует входные аудиоданные. Эта информация воспринимается как имеющая такие характеристики, как скорость звука, высота звука и то, как она меняется со временем.

2 Декодер: декодирует закодированную информацию и преобразует этот звук в текст. Основная функция декодера — дать интерпретацию относительно структуры языка и подлинности слов.

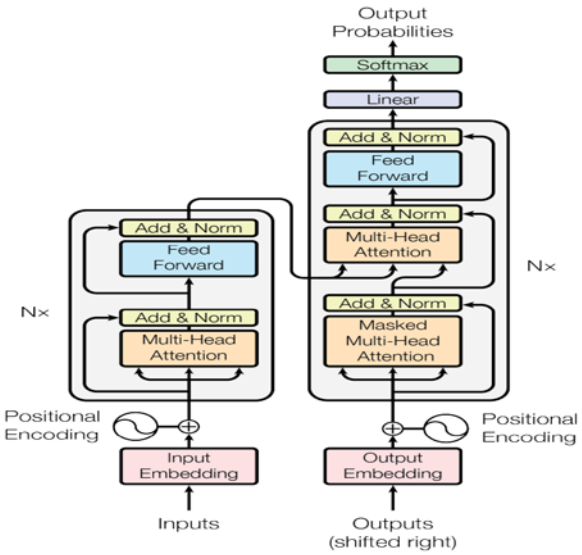

Рисунок 1. Как работают Transformer

Подготовка вашего голоса для индивидуальных моделей: Модель генерирует спектрограммы звуков, используя входные данные. Спектрограммы — это графическое представление звуковой информации, разделенной по частоте и времени. Эти данные служат исходными входными данными для кодировщика.

Технические специалисты по архитектуре:

1 Использование спектрограммы. Входной аудиофайл (MP3, WAV) преобразуется в спектрограмму. Эти данные обрабатываются с частотой 16 кГц. Архитектура модели позволяет быстро и с высокой точностью улавливать основную информацию о звуке при использовании спектрограммы.

2 Слои трансформатора. Число слоев преобразователя в кодере и декодере составляет от 12 до 24, что позволяет выполнять многослойные преобразования. Каждый слой:

- проверяет связь между данными;
- преобразует звуки в текст с учетом структуры языка.

Размеры на разных компьютерах. Производительность модели различается на разных компьютерных платформах. Ключевую роль играют такие характеристики, как мощный процессор, память графического процессора и объем оперативной памяти используемой платформы.

Тесты и измерения. Работоспособность модели проверена в различных конфигурациях. Результаты этих испытаний показаны ниже:

Таблица 1. Работоспособность платформ

Платформа	**Время обработки**	**Использование**	**Использование**

	данных	**графического процессора**	**оперативной памяти**
Intel i5, 8GB RAM	50 **секунд**	нет	4 GB
NVIDIA RTX 3060, 32GB RAM	2 секунд	да	5 GB

Анализ:

1 Стандартные процессоры (ЦП): модель работает с процессорами Intel и AMD, но время работы без ГП увеличивается.

2 При использовании графического процессора: модель работает очень быстро с графическими процессорами, такими как NVIDIA RTX 3060. Генерация спектрограммы и декодирование текста ускоряются с помощью графического процессора.

3 Влияние памяти (ОЗУ): устройства с большим объемом оперативной памяти повышают способность модели обрабатывать большие аудиофайлы.

Практическое использование архитектуры: Модель может использоваться во многих приложениях:

1 Офлайн-перевод: предоставление услуг перевода для людей, говорящих на разных языках.

2 Анализ голоса: позволяет оператору проверить громкость и использование голоса.

3 Регулирование информации: Управление форматом аудиозаписей.

Оптимизация Теперь давайте протестируем модель, изменив некоторые параметры на этапе оптимизации.

Первоначально размер пакета каждого сеанса обучения составлял 24, теперь: мы изменили размер пакета до 16, скорость обучения ранее была 0,001, теперь мы изменили ее на 1e-5, количество шагов обучения было 500, теперь мы увеличили его до 1000, и мы перешли на технологию токенизации OpenAI. Эти обновления оказались особенно важными для повышения точности обучения модели.

Методология: Размер партии: установите 16. Такой размер способствует обработке данных и эффективному использованию памяти графического процессора. Скорость обучения: использовалась низкая скорость 1e-5. Такая скорость снижает опасные колебания при обучении модели с более сложными данными и обеспечивает стабильную сходимость. Шаги: количество шагов увеличено до 1000. Это позволило модели взаимодействовать с большим количеством данных, что позволило ей достичь эффективного уровня обучения.

Токенизатор: переключен на систему токенизаторов OpenAI. Этот токенизатор характеризуется лучшим сохранением контекста данных.

Поскольку архитектура Lexical Analysis Transformer предназначена для обработки прямых цифровых данных, между текстом и токенами должен быть процесс перевода. Токен – целое число, представляющее символ или короткий сегмент символов. На входной стороне входной текст разбивается на последовательность токенов. Между тем, на стороне вывода выходные токены преобразуются обратно в текст. Модуль, который преобразует текст в последовательности токенов, называется «токенизатор».

Все токены образуют словарную базу токенизатора, а ее размер называется размером словаря. Для токенов, находящихся за пределами словаря, обычно используется специальный токен, который часто записывается как «[UNK]» (неизвестно).

Встраивание слов Каждый токен преобразуется в вектор встраивания путем поиска его в специальной таблице. Другими словами, однократное представление токена умножается на матрицу вложения M. Например, если токен доступа равен 3, то горячая форма будет иметь вид:

[0, 0, 0, 1, 0, 0, …]

Вектор внедрения становится следующим:

Встроить(3) = [0, 0, 0, 1, 0, 0, …] M

Векторы внедрения токенов добавляются к соответствующим им векторам кодирования позиций, и они образуют последовательность входных векторов.

Число измерений векторов внедрения называется «скрытым размером» или «измерением внедрения».

Отсоединение Слой извлечения представляет собой процесс, обратный слою внедрения. Слой внедрения преобразует токен в вектор, а слой извлечения преобразует вектор в распределение вероятностей между токенами.

Слой деинсталляции состоит из линейно-мягкого слоя:

UnEmbed(x) = softmax(xW + b)

Заключение Архитектура обеспечивает высокоуровневые возможности для эффективного, быстрого и точного преобразования аудиоданных в текст. Эта система в сочетании с моделью трансформатора может быть полезна не только для технологии распознавания голоса, но и в системах изучения языка и информационных системах.

При тестировании производительности Whisper на различных компьютерных системах мы обнаружили, что лучшие результаты достигаются при использовании графических процессоров. Это показывает важность графических процессоров в отрасли информационных технологий.

Для того чтобы архитектура, используемая для преобразования речи в текст, была успешной, особое внимание необходимо уделить конфигурации технологии.

Список литературы

[1] Чарыев А, Информационные системы и технологии Научное издательство «Символ науки» 2024 13-16 с

[2] Мурадов М Назарова С, Современные технологии распознавания речи Научное издательство «Наука и мировоззрение» 2024 301-305 с

[3] Мурадов М., Распознавание речи Сборник статей Международной научно-практической конференции Научный потенциал - 2023 Петрозаводск 2023 405-409 с

Авторский вклад

Мырадов Максат Тачмухаммедович - авторский вклад в архитектуру Transformer для преобразования голоса в текст, разработав, например, новый метод адаптации механизма внимания, оптимизирующий процесс обучения или интегрировав инновационные слои для улучшения обработки аудиоданных и повышения точности распознавания речи.

TRANSFORMER ARCHITECTURE FOR VOICE-TO-TEXT CONVERSION

M. T. Myradov
Head of the Department of
Information Systems,
The Institute of Telecommunications and
Informatics

Abstract: The Transformer architecture is successfully applied to speech-to-text conversion due to its ability to efficiently process data sequences. The model uses self-attention mechanisms to identify dependencies between input data elements, which improves speech recognition accuracy. This approach allows the system to be trained on large volumes of audio data and achieve high performance in real-world applications.

Keywords: Transformer architecture, speech-to-text conversion, attention, speech recognition.