Архитектура Transformer

Тестовое задание

Архитектура трансформеров стала основой для множества современных моделей в области обработки естественного языка и других задач, связанных с последовательными данными. Основная работа, представившая трансформеры, была опубликована в 2017 году под названием "Attention is All You Need".

Принцип работы и особенности архитектуры

Механизм внутреннего внимания

Фундаментальная операция, выполняемая в реализации любой архитектуры трансформера, представлена механизмом внутреннего внимания (self-attention). Механизм внимания позволяет модели фокусироваться на определенных частях входной последовательности. Это важно, потому что не все слова или элементы входа имеют одинаковую значимость для задачи. Кроме того, трансформеры фокусируются на связных частях предложения, то есть оно ищет те слова, которые относятся друг к другу. Давайте приведем пример: «The animal didn't cross the street because it was too tired». В данном предложении нужно понять, к чему относится слово it – к animal или street (рис. 1). Такую проблему также решает механизм внутреннего внимания.

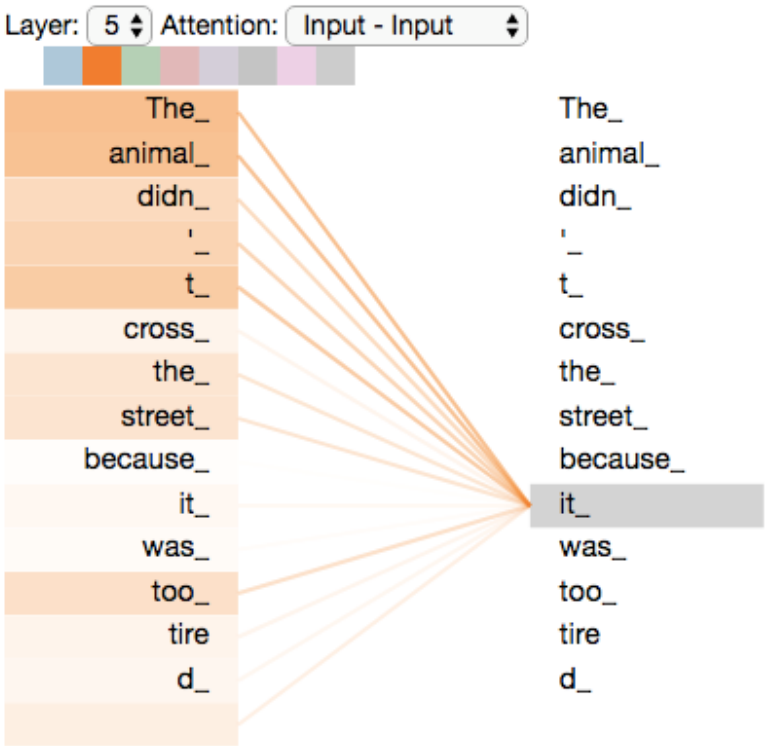

Рисунок 1

В трансформерах используется механизм "Scaled Dot-Product Attention", который вычисляет взвешенные суммы векторов на основе их сходства. В

реализации данного механизма имеются три матрицы: Q, K, V, которые соотносятся следующим образом:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V$$

- Q (Query) — вектор запроса.
- K (Key) — вектор ключа.
- V (Value) — вектор значения.
- d — размерность ключей, используемая для масштабирования.

Если говорить о смысле данных матриц, то на простом языке это можно написать так: Query - насколько каждое слово полезно данному, Key - насколько данное слово полезно другим, Value - чем оно полезно другим (то есть информация, которую оно несет).

Многоголовое внимание

Вместо одного механизма внимания трансформеры используют несколько "голов" внимания, что позволяет модели захватывать различные представления данных. Каждая голова выполняет свою версию внимания, а затем их результаты конкатенируются и проходят через линейный слой.

Архитектура Transformer

Трансформер состоит из двух основных частей: кодировщика (encoder) и декодировщика (decoder).

a. Кодировщик

- Состоит из нескольких слоев (обычно 6) (рис. 2).
- Каждый слой включает в себя многоголовое внимание и полносвязную сеть.
- Входные данные кодировщика представляют собой последовательность векторов, которые могут быть получены из векторных представлений слов (например, с помощью эмбеддингов).
- Использует позиционное кодирование для учета порядка слов в последовательности, так как механизм внимания не имеет информации о последовательности.

b. Декодировщик

- Также состоит из нескольких слоев, аналогичных кодировщику, но с дополнительным механизмом внимания, который позволяет декодировщику обращать внимание на выходы кодировщика (рис. 3).

- На каждом этапе декодирования используется маскированное внимание, чтобы предотвратить утечку информации из будущих токенов (то есть предсказывать только следующее слово на основе предыдущих).

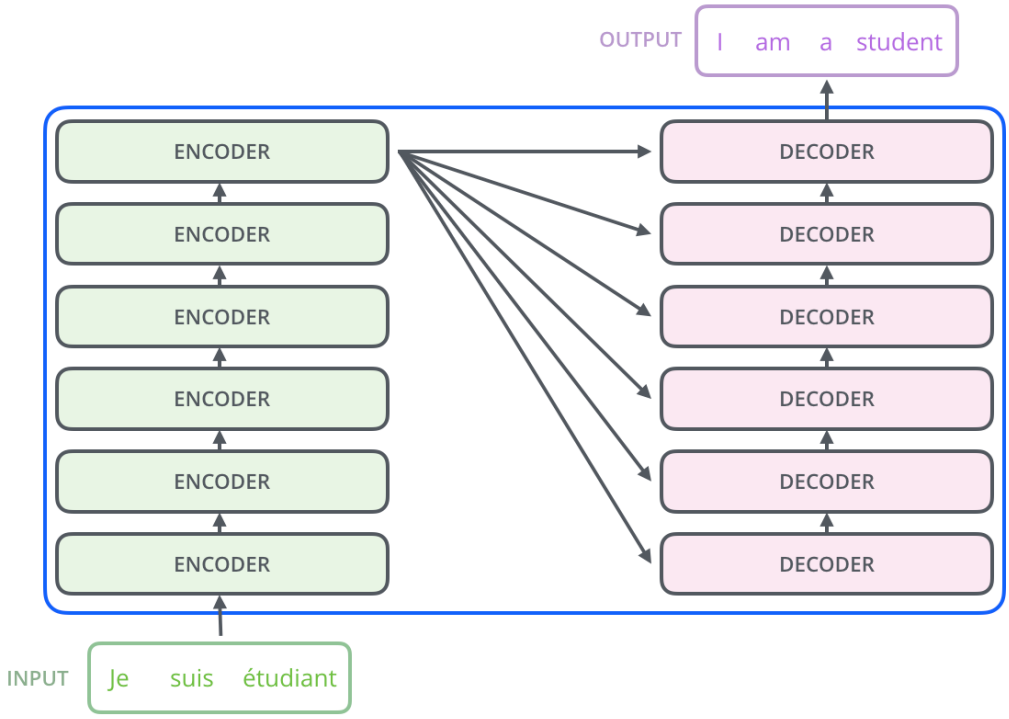

Рисунок 2

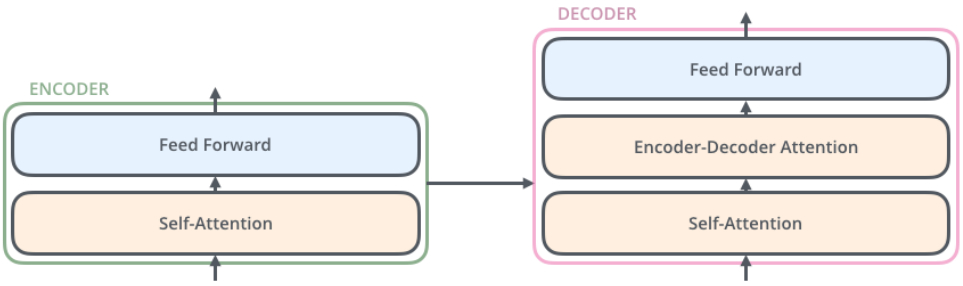

Рисунок 3

Поскольку трансформеры не имеют встроенного представления последовательности, как, например, рекуррентные нейронные сети (RNN), используется позиционное кодирование. Оно добавляет информацию о позиции токенов в последовательности. Позиционное кодирование обычно основано на синусоидальных функциях:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right)$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d}}\right)$$

где pos — позиция токена в последовательности, а d — размерность векторов представления.

Применение архитектуры в задачах глубокого обучения

Обработка естественного языка (NLP)

а) Перевод текста

Пример применения: Модель Transformer из оригинальной статьи «Attention is All You Need» была разработана именно для машинного перевода, например перевода с английского на немецкий и наоборот.

Успешность: Трансформеры обеспечили значительное улучшение качества перевода по сравнению с предыдущими подходами, такими как RNN и LSTM. Они стали стандартом для систем машинного перевода, например, Google Translate использует их для своих услуг.

б) Генерация текста

Пример применения: Модели, такие как GPT (Generative Pre-trained Transformer), используют трансформеры для генерации связного текста на основе заданного контекста.

Успешность: Эти модели способны генерировать высококачественные, осмысленные тексты и используются в приложениях, таких как автоматическая генерация статей, чат-боты и инструменты для креативного письма.

в) Анализ настроений

Пример применения: Трансформеры, такие как BERT (Bidirectional Encoder Representations from Transformers), применяются для анализа настроений в отзывах, социальных медиа и других текстах.

Успешность: BERT значительно улучшил результаты на различных задачах анализа настроений, достигая новых рекордов на стандартных датасетах, таких как GLUE.

Классификация текста

Пример применения: Трансформеры могут использоваться для классификации текстов, таких как классификация новостей по категориям или определение жанра.

Успешность: Модели на основе трансформеров, такие как RoBERTa, продемонстрировали выдающиеся результаты в классификации, значительно превышая точность более традиционных методов, таких как SVM и Naive Bayes.

Ответ на вопросы

Пример применения: Трансформеры применяются в системах, предназначенных для ответов на вопросы, таких как SQuAD (Stanford Question Answering Dataset).

Успешность: Модели, такие как ALBERT и T5, показывают высокую точность в нахождении ответов на вопросы, что делает их эффективными для построения систем вопрос-ответ.

Резюме и аннотирование текста

Пример применения**:** Модели на основе трансформеров могут использоваться для автоматического создания резюме документов и статей.

Успешность: Модели, такие как BART и T5, достигли значительных успехов в этой области, обеспечивая краткие и содержательные аннотации.

Обработка изображений (компьютерное зрение)

Пример применения: Архитектура Vision Transformer (ViT) применяет идеи трансформеров к задачам компьютерного зрения, таким как классификация изображений.

Успешность: ViT достигли сопоставимых, а в некоторых случаях даже превосходящих результатов по сравнению с традиционными свёрточными нейронными сетями (CNN), особенно на больших наборах данных.

Актуальность

Трансформеры по-прежнему являются основой современных решений в различных областях глубокого обучения. Наиболее известными для нас являются GPT-3 и GPT-4. Тем не менее, развитие новых архитектур и подходов, таких, продолжает происходить, предлагая дополнительные возможности и решения, направленные на улучшение производительности, снижение затрат на вычисления и повышение эффективности обучения, например M2M-100 для перевода текстов, RAG для генерации текста и Swin Transformer для компьютерного зрения.