**Лекция 9**

**Трансформеры**

Трансформеры — относительно новый тип нейросетей, направленный на решение последовательностей с легкой обработкой дальнодействующих зависимостей. На сегодня это самая продвинутая техника в области обработки естественной речи (NLP).

С их помощью можно переводить текст, писать стихи и статьи и даже генерировать компьютерный код.В отличие от рекуррентных нейронных сетей (RNN), трансформеры не обрабатывают последовательности по порядку. Например, если исходные данные — текст, то им не нужно обрабатывать конец предложения после обработки начала. Благодаря этому такую нейросеть можно распараллелить и обучить значительно быстрее.

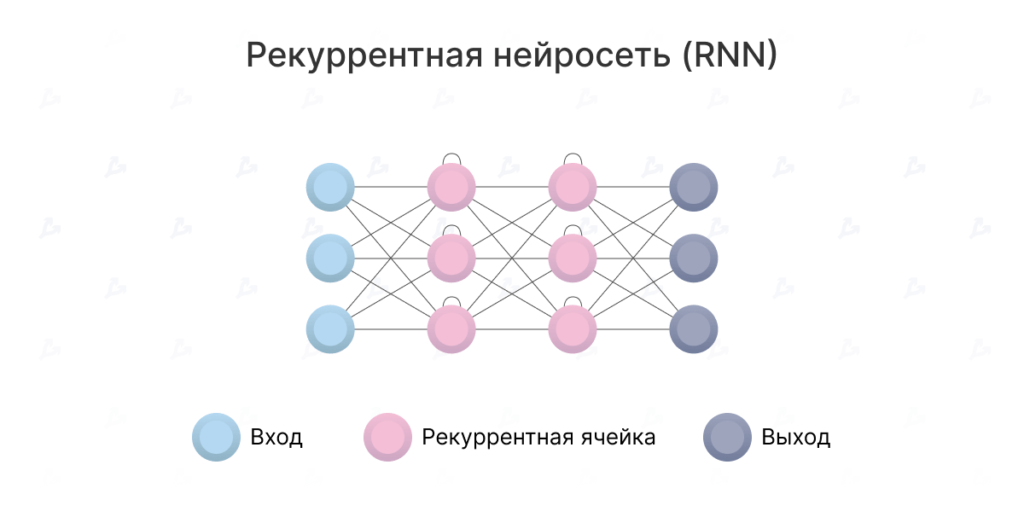
**Когда они появились?**

Трансформеры впервые описали инженеры из Google Brain в работе «Attention Is All You Need» в 2017 году.

Одно из основных отличий от существующих методов обработки данных заключается в том, что входная последовательность может передаваться параллельно, чтобы можно было эффективно использовать графический процессор, а также увеличивать скорость обучения.

**Зачем нужны трансформеры?**

До 2017 года инженеры использовали глубокое обучение для понимания текста с помощью рекуррентных нейронных сетей.



Допустим, при переводе предложения с английского на русский RNN будет принимать в качестве входных данных английское предложение, обрабатывать слова по одному, а затем последовательно выдавать их русские аналоги. Ключевое слово здесь — «последовательный». В языке важен порядок слов, и их нельзя просто перемешать.

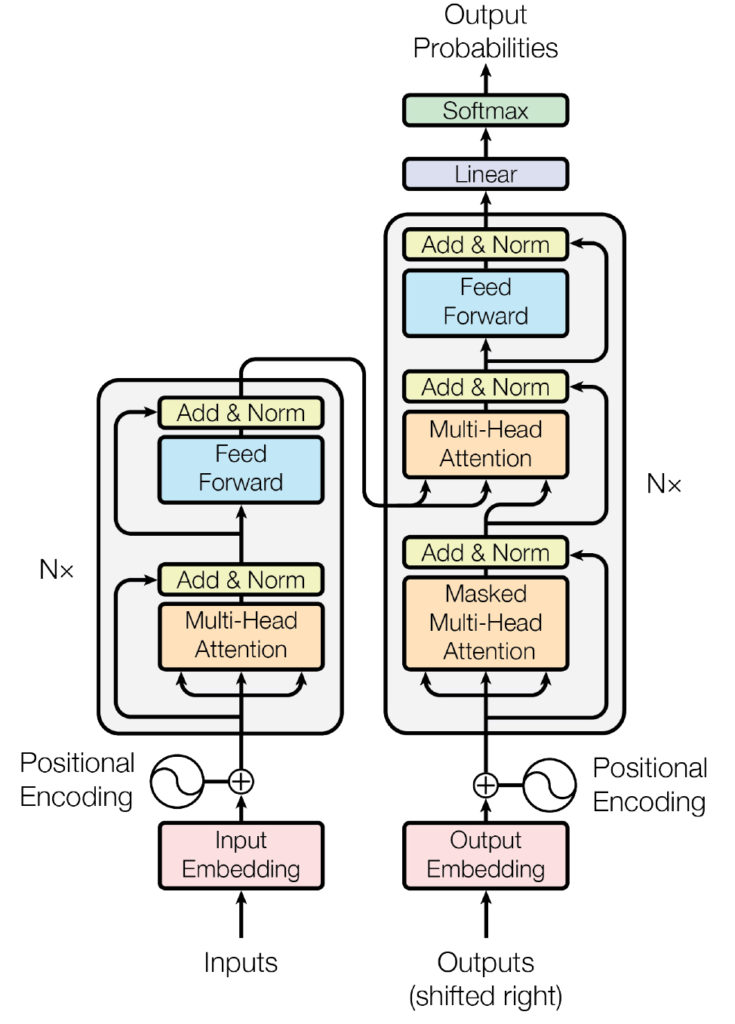
Здесь RNN сталкиваются с рядом проблем. Во-первых, они пытаются обрабатывать большие последовательности текста. К моменту продвижения к концу абзаца они «забывают» содержание начала. Например, модель перевода на основе RNN может иметь проблемы с запоминанием пола объекта длинного текста.

Во-вторых, RNN сложно тренировать. Как известно, они подвержены так называемой проблеме исчезающего/взрывающегося градиента.

В-третьих, они обрабатывают слова последовательно, рекуррентную нейросеть трудно распараллелить. Это значит, что ускорить обучение, используя больше графических процессоров невозможно. Следовательно, что ее нельзя обучить на большом количестве данных.

**Как работают трансформеры?**

Основными компонентами трансформеров являются энкодер и декодер.



Архитектура трансформеров. Данные: работа [«Attention Is All You Need»](https://arxiv.org/abs/1706.03762).

Энкодер преобразовывает входящую информацию (например, текст) и конвертирует ее в вектор (набор чисел). Декодер, в свою очередь, расшифровывает ее в виде новой последовательности (например, ответ на вопрос) слов на другом языке — смотря для каких целей создавалась нейросеть.

Другие инновации, лежащие в основе трансформеров, сводятся к трем основным концепциям:

* позиционные энкодеры (Positional Encodings);
* внимание (Attention);
* самовнимание (Self-Attention).

Начнем с первого — позиционных энкодеров. Допустим, необходимо перевести текст с английского на русский. Стандартные модели RNN «понимают» порядок слов и обрабатывают их последовательно. Однако это затрудняет распараллеливание процесса.

Позиционные кодировщики позволяют преодолеть этот барьер. Идея состоит в том, чтобы взять все слова во входной последовательности — в данном случае английское предложение — и добавить к каждому номер в его порядке. Итак, вы «скармливаете» сети такую ​​последовательность:

**[(“Red”, 1), (“fox”, 2), (“jumps”, 3), (“over”, 4), (“lazy”, 5), (“dog”, 6)]**

Концептуально это можно рассматривать как перенос бремени понимания порядка слов со структуры нейронной сети на сами данные.

Сначала, прежде чем трансформеры обучатся на какой-либо информации, они не знают, как интерпретировать эти позиционные кодировки. Но по мере того как модель видит все больше и больше примеров предложений и их кодировок, она учится эффективно их использовать.

Представленная выше структура приведена в чрезмерно упрощенном виде — авторы оригинального исследования использовали синусоидальные функции, чтобы придумать позиционные кодировки, а не простые целые числа 1, 2, 3, 4, но суть та же. Сохраняя порядок слов в виде данных, а не структуры, нейронную сеть легче обучать.

Внимание — это структура нейронной сети, [введенная в контекст машинного перевода в 2015 году](https://arxiv.org/abs/1409.0473). Чтобы понять эту концепцию, обратимся к оригинальной статье.

Представим, что нам нужно перевести на французский язык фразу:

**«The agreement on the European Economic Area was signed in August 1992».**

Французский эквивалент выражения звучит следующим образом:

**«L’accord sur la zone économique européenne a été signé en août 1992».**

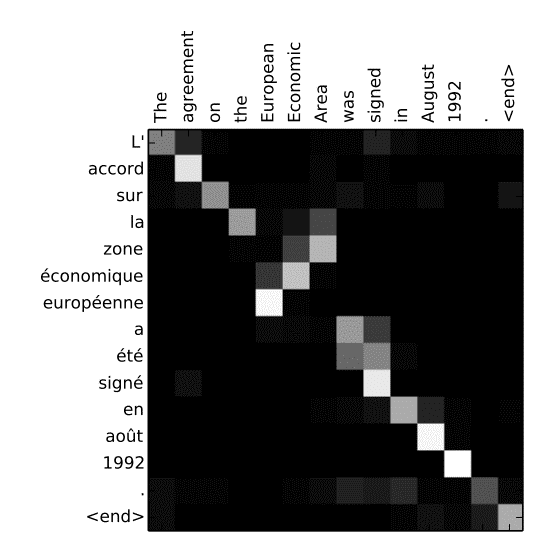
Наихудшим вариантом перевода является прямой поиск аналогов слов из английского языка во французском, одного за другим. Этого не получится сделать по нескольким причинам.

Во-первых, некоторые слова во французском переводе перевернуты:

**«European Economic Area»** против **«la zone économique européenne»**.

Во-вторых, французский язык богат гендерными словами. Чтобы соответствовать женскому объекту **«la zone»**, прилагательные **«économique»** и **«européenne»** также необходимо поставить в женский род.

Внимание помогает избегать таких ситуаций. Его механизм позволяет текстовой модели «смотреть» на каждое слово в исходном предложении при принятии решения о том, как их переводить. Это демонстрирует визуализация из оригинальной статьи:



Визуализация машинного перевода с использованием механизмов внимания. Данные: «Neural Machine Translation by Jointly Learning to Align and Translate (2015)».

Это своего рода тепловая карта, показывающая, на что модель «обращает внимание», когда переводит каждое слово во французском предложении. Как и следовало ожидать, когда модель выводит слово **«européenne»**, она в значительной степени учитывает оба входных слова — **«European»** и **«Economic»**.

Узнать модели, на какие слова «обращать внимание» на каждом шаге, помогают тренировочные данные. Наблюдая за тысячами английских и французских предложений, алгоритм узнает взаимозависимые типы слов. Он учится учитывать пол, множественность и другие правила грамматики.

Механизм внимания был чрезвычайно полезным инструментом для обработки естественного языка с момента его открытия в 2015 году, но в своей первоначальной форме он использовался вместе с рекуррентными нейронными сетями. Таким образом, нововведение статьи 2017 года о трансформерах было частично направлено на то, чтобы полностью отказаться от RNN. Вот почему работа 2017 года называется «Внимание — это все, что вам нужно».

Последняя часть трансформеров — это поворот внимания, называемый «самовниманием».

Если внимание помогает выравнивать слова при переводе с одного языка на другой, то самовнимание позволяет модели понимать смысл и закономерности языка.

Например, рассмотрим эти два предложения:

**«Николай потерял ключ от машины»**

**«Журавлиный ключ направился на юг»**

Слово **«ключ»** здесь означает две очень разные вещи, которые мы, люди, зная ситуацию, можем запросто отличать их значения. Самовнимание позволяет нейронной сети понимать слово в контексте слов вокруг него.

Поэтому, когда модель обрабатывает слово **«ключ»** в первом предложении, она может обратить внимание на **«машины»** и понять, что речь идет о металлическом стержне особой формы для замка, а не что-то другое.

Во втором предложении модель может обратить внимание на слова **«журавлиный»** и **«юг»**, чтобы отнести **«ключ»** к стае птиц. Самовнимание помогает нейронным сетям устранять неоднозначность слов, делать частеречную разметку, изучать семантические роли и многое другое.