



Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет
имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ

Робототехника и комплексная автоматизация (РК)

КАФЕДРА

Системы автоматизированного проектирования (РК6)

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА
К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ
НА ТЕМУ:

«Сравнение языковых моделей GPT3, GPT3.5 и GPT4»

Студент РК6-81Б

(Подпись, дата)

Гунько Н.М.

И.О. Фамилия

Руководитель

(Подпись, дата)

Витюков Ф.А.

И.О. Фамилия

Москва, 2023 г.

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

УТВЕРЖДАЮ
Заведующий кафедрой РК6
А.П. Карпенко

«_____» _____ 2023 г.

ЗАДАНИЕ
на выполнение научно-исследовательской работы

по теме: Сравнение языковых моделей GPT3, GPT3.5 и GPT4

Студент группы РК6-81Б

Гунько Никита Макарович
(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.) учебная
Источник тематики (кафедра, предприятие, НИР) предприятие

График выполнения НИР: 25% к 5 нед., 50% к 11 нед., 75% к 14 нед., 100% к 16 нед.

Техническое задание: Исследовать понятие языковой модели и описать ее архитектуру, рассмотреть алгоритм работы и описать устройство модели GPT и выполнить сравнение версий 3, 3.5 и 4.

Оформление научно-исследовательской работы:

Расчетно-пояснительная записка на 15 листах формата А4.

Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.):

Дата выдачи задания «15» мая 2023 г.

Руководитель НИР

(Подпись, дата)

Витюков Ф.А.

И.О. Фамилия

Студент

(Подпись, дата)

Гунько Н.М.

И.О. Фамилия

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

СОДЕРЖАНИЕ

| | |
|---|----|
| ВВЕДЕНИЕ | 4 |
| 1. Языковое моделирование | 5 |
| 1.1. Обзор концепций..... | 5 |
| 1.2. Классификация языковых моделей | 5 |
| 1.3. Возможности языковых моделей | 7 |
| 2. Языковая модель GPT..... | 9 |
| 2.1. Общие сведения и архитектура | 9 |
| 2.2. GPT-1 | 10 |
| 2.3. GPT-2..... | 11 |
| 2.4. GPT-3..... | 12 |
| 2.5. GPT-4..... | 13 |
| ЗАКЛЮЧЕНИЕ | 14 |
| СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ | 15 |

ВВЕДЕНИЕ

В последние годы ученые и исследователи в области обработки естественного языка все больше обращают внимание на языковую модель GPT. Эта модель искусственного интеллекта стала широко известной и приобрела значительную популярность. Она способна генерировать текст, который кажется похожим на то, что может написать человек, и демонстрирует более сложное понимание естественного языка.

Языковые модели GPT (Generative Pre-trained Transformer) разработаны компанией OpenAI и представляют собой одни из самых мощных моделей искусственного интеллекта для обработки естественного языка. Они обладают способностью генерировать качественные тексты, а также демонстрируют значительное понимание контекста и смысла языковых конструкций.

Сравнение различных версий языковых моделей GPT позволит оценить их преимущества, улучшения и потенциал в области обработки естественного языка, а также покажет исследователям, разработчикам и пользователям как более эффективно применять эти модели в своих проектах и задачах, и еще предоставит важную информацию для определения будущих направлений исследований и разработок в области языковых моделей.

Цель работы: исследовать архитектуру и алгоритм работы языковой модели GPT и выполнить обзор трех ее последних версий.

Для выполнения поставленной цели необходимо решение следующих задач:

- изучить понятие языковой модели;
- рассмотреть виды языковых моделей;
- описать архитектуру языковой модели GPT;
- выполнить сравнение GPT-1, GPT-2, GPT-3-3.5 и GPT4.

1. Языковое моделирование

1.1. Обзор концепций

Языковое моделирование (Language Modeling) – это использование различных статистических и вероятностных методов для определения вероятности появления определенной последовательности слов в предложении. Языковая модель – это тип модели машинного обучения, обученной проводить распределение вероятностей по словам. Проще говоря, модель пытается предсказать следующее наиболее подходящее слово для заполнения пробела в предложении или фразе, исходя из контекста данного текста. Языковые модели анализируют массивы текстовых данных, чтобы обеспечить основу для своих предсказаний слов. Они используются в приложениях обработки естественного языка (NLP), особенно в тех, которые генерируют текст в качестве вывода. Некоторые из этих приложений включают в себя, машинный перевод и ответы на вопросы [1].

1.2. Классификация языковых моделей

Существует несколько различных вероятностных подходов к моделированию языка, которые различаются в зависимости от назначения языковой модели. С технической точки зрения различные типы различаются объемом текстовых данных, которые они анализируют, и математическими расчетами, которые они используют для их анализа. Например, языковая модель, предназначенная для генерации предложений для автоматизированного бота, может использовать другую математику и анализировать текстовые данные иначе, чем языковая модель, предназначенная для определения вероятности поискового запроса.

Некоторые распространенные типы статистических языковых моделей включают:

- N-граммы. N-граммы представляют собой относительно простой подход

к языковым моделям. Они создают вероятностное распределение для последовательности из n элементов. Значение n может быть любым числом и определяет размер “грамма” или последовательности слов, которой присваивается вероятность. Например, если $n = 5$, один элемент N -граммы может выглядеть так: “вы можете позвонить мне, пожалуйста”. Затем модель присваивает вероятности, используя последовательности размера n . В основном, n можно рассматривать как количество контекста, которое модель должна учитывать. Частными видами n -граммов являются униграммы, биграммы, триграммы и так далее.

- Униграмма. Униграмма является самым простым типом языковой модели. Она не учитывает контекст при вычислениях. Она оценивает каждое слово или термин независимо. Униграммы часто используются для задач обработки языка, таких как информационный поиск. Униграмма является основой для более специфической модели, называемой моделью вероятности запроса, которая использует информационный поиск для анализа набора документов и подбора наиболее релевантного для конкретного запроса.

- Двунаправленная. В отличие от n -граммов, которые анализируют текст в одном направлении (назад), двунаправленные модели анализируют текст в обоих направлениях, вперед и назад. Эти модели могут предсказывать любое слово в предложении или тексте, используя каждое другое слово в тексте. Исследование текста в обоих направлениях повышает точность результатов. Этот тип модели часто используется в приложениях машинного обучения и генерации речи. Например, Google использует двунаправленную модель для обработки поисковых запросов.

- Экспоненциальная. Также известные как модели максимальной энтропии, этот тип является более сложным, чем n -граммы. Проще говоря, модель оценивает текст с помощью уравнения, которое объединяет функции признаков и n -граммы. В основном, этот тип указывает функции и параметры желаемых результатов и, в отличие от n -грамм, оставляет параметры анализа более неопределенными – например, он не указывает размеры отдельных

граммов. Модель основана на принципе энтропии, который гласит, что наиболее хаотичное вероятностное распределение является наилучшим выбором. Другими словами, модель с наибольшим хаосом и меньшим количеством предположений является наиболее точной. Экспоненциальные модели разработаны для максимизации перекрестной энтропии, что минимизирует количество статистических предположений, которые можно сделать. Это позволяет пользователям больше доверять полученным от этих моделей результатам

- Непрерывное пространство. Этот тип модели представляет слова как нелинейную комбинацию весов в нейронной сети. Процесс присвоения веса слову также известен как вложение слова. Этот тип модели становится особенно полезным с увеличением размера набора данных, поскольку в больших наборах данных часто присутствуют больше уникальных слов.

Присутствие большого количества уникальных или редко используемых слов может вызвать проблемы для линейных моделей, таких как n-граммы. Это связано с тем, что количество возможных последовательностей слов увеличивается, и шаблоны, определяющие результаты, становятся слабее. Путем взвешивания слов нелинейным и распределенным образом эта модель может “учиться” приближать слова и, следовательно, не поддаваться влиянию неизвестных значений. Ее “понимание” данного слова не так тесно связано с непосредственно окружающими словами, как в моделях n-грамм. Наиболее часто используемые архитектуры нейронных сетей для задач NLP – это рекуррентные нейронные сети (RNN) и сети-трансформеры (Transformers).

1.3. Возможности языковых моделей

Языковые модели используются в различных задачах NLP, таких как распознавание речи, машинный перевод и обобщение текста. Некоторые распространенные задачи, в которых используются языковые модели:

- Генерация контента. Одной из областей, в которой языковые модели

проявляют себя ярче всего, является создание контента. Это включает в себя создание полных текстов или их частей на основе данных и терминов, предоставленных людьми. Контент может варьироваться от новостных статей, пресс-релизов и сообщений в блогах до описаний продуктов интернет-магазина и стихов и это лишь некоторые из них.

- Маркировка части речи (POS). Языковые модели широко используются для достижения самых современных результатов в задачах тегирования POS. Тегирование POS – это процесс маркировки каждого слова в тексте соответствующей частью речи, такой как существительное, глагол, прилагательное и т. д. Модели обучаются на большом объеме размеченных текстовых данных и могут научиться предсказывать POS слова. на основе его контекста и окружающих слов в предложении.

- Ответ на вопрос. Языковые модели можно научить понимать вопросы и отвечать на них в заданном контексте и без него. Они могут давать ответы несколькими способами, например, извлекая определенные фразы, перефразируя ответ или выбирая из списка вариантов.

- Обобщение текста. Языковые модели можно использовать для автоматического сокращения документов, статей, подкастов, видео и многого другого до наиболее важных фрагментов. Модели могут работать двумя способами: извлекать наиболее важную информацию из исходного текста или предоставлять резюме, которые не повторяют исходный язык.

- Анализ настроений. Подход языкового моделирования является хорошим вариантом для задач анализа настроений, поскольку он может уловить тон голоса и семантическую ориентацию текстов.

- Разговорный ИИ. Языковые модели – неизбежная часть речевых приложений, требующих преобразования речи в текст и наоборот. Являясь частью диалоговых систем ИИ, языковые модели могут предоставлять соответствующие текстовые ответы на входные данные.

- Машинный перевод. Способность языковых моделей на основе машин-

ного обучения эффективно обобщать длинные контексты позволила им улучшить машинный перевод. Вместо того, чтобы переводить текст слово за словом, языковые модели могут изучать представления входных и выходных последовательностей и обеспечивать надежные результаты.

- Завершение кода. Последние крупномасштабные языковые модели продемонстрировали впечатляющую способность генерировать, редактировать и объяснять код. Однако они могут выполнять только простые задачи программирования, переводя инструкции в код или проверяя его на наличие ошибок.

Это всего лишь несколько вариантов использования языковых моделей: их потенциал гораздо значительнее [2].

2. Языковая модель GPT

2.1. Общие сведения и архитектура

GPT расшифровывается как генеративный предварительно-обученный трансформер (GPT) – это тип модели машинного обучения, используемой для задач обработки естественного языка. Эти модели предварительно обучаются на больших объемах данных, таких как книги и веб-страницы, для создания контекстуально релевантного и семантически связного языка.

В основе языковой модели GPT лежит несколько блоков нейронной сети архитектуры Transformer.

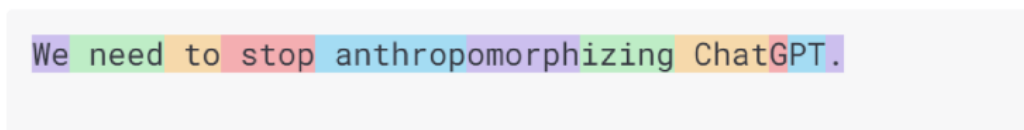
Начнем с того, что генеративные модели получают в качестве входных данных какое-то количество токенов, и создают один токен в качестве выходных данных (рис. 1).



Рисунок 1 – Входные и выходные данные генеративной модели

Это кажется довольно простой концепцией, но, чтобы понять ее, нам нужно знать, что такое токен. Токен – это фрагмент текста. В контексте моделей GPT компании OpenAI общие и короткие слова обычно соответствуют одному токenu, например, слово “Мы” на рисунке 2. Длинные и редко используемые слова обычно разбиваются на несколько токенов. Например, слово “антропоморфизация” на рисунке 2 разбито на три токена. Аббревиатуры, такие как “ChatGPT”, могут быть представлены одним токеном или разбиты на несколько, в зависимости от того, насколько часто буквы появляются вместе.

| Tokens | Characters |
|--------|------------|
| 11 | 43 |



We need to stop anthropomorphizing ChatGPT.

Рисунок 2 – Пример разделения на токены

2.2. GPT-1

GPT-1 был выпущен компанией OpenAI в 2018 году как их первая итерация языковой модели с использованием архитектуры Transformer. У него было 117 миллионов параметров, что значительно улучшило предыдущие современные языковые модели.

Одной из сильных сторон GPT-1 была его способность генерировать плавный и связный язык при наличии подсказки или контекста. Модель была обучена на сочетании двух наборов данных: Common Crawl, массивного набора данных веб-страниц с миллиардами слов, и набора данных BookCorpus, коллекции из более чем 11 000 книг различных жанров. Использование этих

разнообразных наборов данных позволило GPT-1 развить сильные способности языкового моделирования.

Хотя GPT-1 был значительным достижением в области обработки естественного языка (NLP), у него были определенные ограничения. Например, модель была склонна генерировать повторяющийся текст, особенно когда ей давали подсказки, выходящие за рамки ее обучающих данных. Он также не мог рассуждать о нескольких оборотах диалога и не мог отслеживать долгосрочные зависимости в тексте. Кроме того, его связность и беглость были ограничены только более короткими текстовыми последовательностями, а более длинным отрывкам не хватало связности.

Несмотря на эти ограничения, GPT-1 заложил основу для более крупных и мощных моделей, основанных на архитектуре Transformer [3].

2.3. GPT-2

GPT-2 был выпущен компанией OpenAI в 2019 году как преемник GPT-1. Он содержал ошеломляющие 1,5 миллиарда параметров, что значительно больше, чем у GPT-1. Модель была обучена на гораздо большем и разнообразном наборе данных, сочетающем Common Crawl и WebText.

Одной из сильных сторон GPT-2 была его способность генерировать связные и реалистичные последовательности текста. Кроме того, он может генерировать ответы, подобные человеческим, что делает его ценным инструментом для различных задач обработки естественного языка, таких как создание контента и перевод.

Однако GPT-2 не лишен недостатков. Он боролся с задачами, которые требовали более сложных рассуждений и понимания контекста. В то время как GPT-2 преуспел в коротких абзацах и фрагментах текста, он не смог сохранить контекст и связность в более длинных отрывках.

Эти ограничения проложили путь к разработке следующей итерации моделей GPT.

2.4. GPT-3

Модели обработки естественного языка совершили экспоненциальный скачок с выпуском GPT-3 в 2020 году. Имея 175 миллиардов параметров, GPT-3 более чем в 100 раз больше, чем GPT-1, и более чем в десять раз больше, чем GPT-2.

GPT-3 обучается на различных источниках данных, включая BookCorpus, Common Crawl и Wikipedia. Наборы данных содержат почти триллион слов, что позволяет GPT-3 генерировать сложные ответы на широкий спектр задач NLP, даже без предоставления каких-либо предварительных данных.

Одним из основных улучшений GPT-3 по сравнению с предыдущими моделями является его способность генерировать связный текст, писать компьютерный код и даже создавать произведения искусства. В отличие от предыдущих моделей, GPT-3 понимает контекст данного текста и может генерировать соответствующие ответы. Возможность создавать естественно звучащий текст имеет огромное значение для таких приложений, как чат-боты, создание контента и языковой перевод. Одним из таких примеров является ChatGPT, диалоговый бот с искусственным интеллектом, который почти за одну ночь превратился из безвестности в известность.

Хотя GPT-3 может делать невероятные вещи, у него все же есть недостатки. Например, модель может возвращать предвзятые, неточные или неуместные ответы. Эта проблема возникает из-за того, что GPT-3 обучается на большом количестве текста, который может содержать предвзятую и неточную информацию. Также бывают случаи, когда модель генерирует совершенно нерелевантный текст для подсказки, что указывает на то, что модель все еще испытывает трудности с пониманием контекста и фоновых знаний.

Возможности GPT-3 также вызвали озабоченность по поводу этических последствий и потенциального неправильного использования таких мощных языковых моделей. Эксперты обеспокоены возможностью использования

модели в злонамеренных целях, таких как создание поддельных новостей, фишинговых писем и вредоносного ПО. Действительно, мы уже видели, как преступники используют ChatGPT для создания вредоносных программ.

OpenAI также выпустила улучшенную версию GPT-3, GPT-3.5, до официального запуска GPT-4.

2.5. GPT-4

GPT-4 – последняя модель в серии GPT, выпущенная 14 марта 2023 года. Это значительный шаг вперед по сравнению с предыдущей моделью GPT-3, которая уже производила впечатление. Хотя особенности обучающих данных и архитектуры модели официально не объявлены, она, безусловно, опирается на сильные стороны GPT-3 и преодолевает некоторые из ее ограничений.

Выдающейся особенностью GPT-4 являются его мультимодальные возможности. Это означает, что модель теперь может принимать изображение в качестве входных данных и понимать его как текстовую подсказку. Например, во время прямой трансляции запуска GPT-4 инженер OpenAI передал модели изображение нарисованного от руки макета веб-сайта, и модель неожиданно предоставила рабочий код для веб-сайта.

Модель также лучше понимает сложные подсказки и демонстрирует производительность на уровне человека в нескольких профессиональных и традиционных тестах. Кроме того, у него больше окно контекста и размер контекста, который относится к данным, которые модель может сохранить в своей памяти во время сеанса чата.

GPT-4 раздвигает границы того, что в настоящее время возможно с помощью инструментов ИИ, и, вероятно, найдет применение в самых разных отраслях. Однако, как и в случае с любой мощной технологией, существуют опасения по поводу потенциального неправильного использования и этических последствий такого мощного инструмента.

ЗАКЛЮЧЕНИЕ

Временно отсутствует.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. *Ben Lutkevich*. Language modeling // TechTarget [Электронный ресурс]. Режим доступа: <https://www.techtarget.com/searchenterpriseai/definition/language-modeling> (дата обращения 20.05.2023);
2. Language Models, Explained: How GPT and Other Models Work // altexsoft [Электронный ресурс]. Режим доступа: <https://www.altexsoft.com/blog/language-models-gpt/> (дата обращения 20.05.2023);
3. *Fawad Ali*. GPT-1 to GPT-4: Each of OpenAI's GPT Models Explained and Compared // MakeUseOf [Электронный ресурс]. Режим доступа: <https://www.makeuseof.com/gpt-models-explained-and-compared/> (дата обращения 21.05.2023);