ОБЗОР МЕТОДОВ ГЕНЕРАЦИИ ТЕКСТА НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Воронова В. В., Удовин И. А.
Кафедра информатики, Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: veronika.voronova31@gmail.com, wilcot@ya.ru

В последние годы были изобретены и развиты многие подходы искусственного интелекта для генерации различных данных: текста, изображений, музыки. Возникает желание понять, могут ли одни и те же методы быть переиспользованы для генерации другого типа данных. Для этого необходимо изучить конкретные методы, применяемые в каждом случае. В данной статье приводятся обзоры подходов, применяемых в области генерации текстов естественного языка, и алгоритмы, которые обеспечивают эти механизмы.

Введение

Одно из значительных достижений человека - способность передавать данные и делиться ими. При обсуждении языка, на котором говорят люди, возникает идея его генерации, ведь язык является одним из наиболее сложных умений человека. По некоторым оценкам, лишь 21 процент[1] существующей информации структурирован. Информация создается повсюду в больших масштабах в виде твитов и сообщений, где большая часть существует в текстовой форме, которая по своей природе очень неструктурирована. Для получения важных знаний из этой информации, важно познакомиться с системами обработки естественного языка(NLP). Обработка естественного языка - область разработки программного обеспечения и искусственного интеллекта, которая работает с человеческими языками. Это подходы, при которых мы автоматически описываем и анализируем язык.

I. Представление данных

Подходы распределенного представления входных данных являются фундаментальной идеей для глубоких генеративных моделей, особенно применительно к задачам обработки естественного языка. Нераспределенное представление делает данные разреженными, что неэффективно по нескольким причинам: размерность данных увеличивается по мере увеличения структуры, а так же повышается глубина генеративных моделей для достижения лучших результатов в задачах. Если набор данных содержит функции, которые имеют схожие значения, лучше получить представление, описывающее эти сходства. Векторное представление слов(word embeddings) - это отображение дискретной категориальной переменной в вектор непрерывных чисел.

Word2Vec - это метод, который предсказывает целевое слово из заданного контекста слов (модель Continuous Bag of Words (CBOW)) или предсказывает контекст из целевого слова (модель Skip-gram)[2]. Проблема Word2Vec в том, что он полагается на локальный контекст предложений, что означает, что он захватывает только семантическую информацию языка. Еще один подход, Glove (Global Vectors), с другой стороны, улавливает как глобальный, так и локальный контекст словарного запаса при преобразовании слов в векторы. У каждого из них есть свои применения: Word2Vec очень хорошо справляется с задачами аналогии, а Glove работает с совпадением слов. Еще один метод представления слов в векторы - FastText - представляет каждое слово как n-граммы символов, а не содержит слова напрямую.

II. Методы, основанные на рекуррентных нейронных сетях

Рекуррентные нейронные сети(RRNs) - мощный алгоритм для задач обработки естественного языка, особенно при моделировании последовательных данных. Поскольку рекуррентные нейронные сети содержат внутреннюю память, благодаря которой они могут запоминать предыдущий ввод, а также текущий ввод, то это значительно упрощает задачи моделирования последовательности. Вывод на любом временном шаге зависит не только от текущего ввода, но и от вывода, сгенерированного на предыдущих временных шагах, что делает алгоритм очень подходящим для таких задач, как генерация языка, языковой перевод, анализ тональности и т. д. Но было замечено, что обучение таких сетей очень сложно, что препятствует их использованию во многих задачах.

Долгая краткосрочная память(LSTM) наследует ту же архитектуру, что и обычные рекуррентные нейронные сети без скрытого состояния. Единицы памяти в LSTM называются ячейками, которые принимают в качестве входных данных комбинацию предыдущего состояния и текущего ввода. Эти ячейки фактически решают, что оставить в памяти, а что удалить. GRU - это еще одно расширение стандартных рекуррентных нейронных сетей, которое изменяет архитектуру LSTM с помощью стробирующей сети, которая генерирует сигналы, которые управ-

ляют текущим вводом и предыдущей памятью, чтобы обновить текущую активацию и текущее состояние сети. Это проще, чем LSTM, в котором обновление параметров также используется для ячеек в соответствии с алгоритмом.

Было предложено множество глубоких генеративных моделей, в которых двунаправленные рекуррентные нейронные сети используется для генерации последовательности выходных данных. Идея таких сетей заключается в том, что выход на временном шаге может зависеть не только от предыдущих элементов, но и от будущих элементов последовательности. Они состоят из двух независимых рекуррентных нейронных сетей.

III. Сверточные нейронные сети

Сверточные нейронные сети - один из популярных алгоритмов, используемых для компьютерного зрения. Их исследования, применяемых для задач обработки естественного языка, были начаты недавно, и дали интересные результаты. Было предложено множество методов классификации текстов с использованием сверточных нейронных сетей. В отличие от задач компьютерного зрения, где в качестве входных данных используются пиксели изображения, в задачах языка вместо пикселей изображения используются предложения, слова или иногда символы, в зависимости от классификационной проблемы.

IV. Вариационные автокодировщики

Популярные модели глубокого обучения требуют большого количества структурированных данных. Маркировка неструктурированных данных занимает очень много времени. Один из способов решения этой проблемы - использовать обучение без учителя(unsupervised) для обучения на данных без меток. Вариационные автокодировщики(Variational Auto-Encoders) - одна из мощных генеративных моделей, работающих с немаркированными данными. Он содержит кодировщик, который кодирует данные в скрытые переменные, а так же декодер, который декодирует эти скрытые переменные для восстановления данных. Вариационные автокодировщики были разработаны как один из популярных способов обучения без учителя для сложных распределений. Их применение для генерации дискретных данных (текста) ограничено. Основная проблема использования для генерации текста - это коллапс KL (это означает, что когда декодер становится более мощным, чем цель обучения, может быть решена с помощью ложной стратегии). Были предложены некоторые методы, чтобы преодолеть проблему, однако они не решили ее полностью.

V. Генеративные состязательные сети

Мы научили машины разбираться в вещах, многие архитектуры глубокого обучения заслуживают признания за их творческий успех. Несмотря на это, многие глубокие генеративные модели не достигли большого успеха из-за их неспособности аппроксимировать трудноразрешимые вероятностные вычисления. Но было найдено решение, которое может обойти эти проблемы, называемое генеративные состязательные сети(Generative Adversarial Networks - GAN). GAN - это популярный алгоритм глубокого обучения, использующий состязательный подход, отличный от обычной нейронной сети. GAN содержит две модели, которые обучаются состязательным способом. Генератор производит выборки данных, дискриминатор - классифицирует эти выборки данных как реальные (обучающие данные) или фальшивые (сгенерированные генератором). Цель генератора - производить выборки, которые очень близки к истинным данным, чтобы можно было обмануть дискриминатор, а цель дискриминатора - точно классифицировать эти два типа выборок данных. Дискриминатор пытается максимизировать целевую функцию, а генератор пытается минимизировать целевую функцию.

VI. Заключение

Для генерации текста естественного языка могут использоваться как методы обучения с учителем, так и без него. Все методы имеют как достоинства, так и слабые места, и чаще всего хороши в решении конкретных задач. Так же необходимо иметь ввиду, что некоторые решения могу быть ограничены вычислительными ресурсами.

Список литературы

1. The survey: Text generation models in deep learning [Электронный ресурс] / sciencedirect.com – Режим доступа: https://www.sciencedirect.com/science/article/pii/S1319157820303360/. – Дата доступа: 18.10.2021.
2. Word2Vec Explained [Электронный ресурс] / towardsdatascience.com – Режим доступа: https://towardsdatascience.com/word2vec-explained-49c52b4ccb71/. – Дата доступа: 18.10.2021.
3. A Gentle Introduction to Generative Adversarial Networks [Электронный ресурс] / machinelearningmastery.com – Режим доступа: https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/. – Дата доступа: 18.10.2021.