

ISSN 2523-4714

УДК 004.89

А. Н. Кузьмин, А. М. Туровец
Институт бизнеса БГУ, Минск, Беларусь

ПРОБЛЕМА МИНИМИЗАЦИИ ОБЩИХ ОШИБОК ЯЗЫКОВЫХ ГЕНЕРАТИВНЫХ МОДЕЛЕЙ НЕЙРОННЫХ СЕТЕЙ В ЛОГИСТИЧЕСКИХ СИСТЕМАХ

Рассмотрена проблема общих ошибок языковых генеративных моделей нейронных сетей, а также предложены пути ее решения. Сформулирована структура входных и выходных данных в результате взаимодействия с языковой генеративной моделью нейронной сети, выделен сектор общих ошибок. Посредством проведения экспериментов на решение задач вероятностного распределения ответов заданной совокупности и заданий со свободным ответом из набора данных, содержащем вопросы по функциональным областям логистики, установлена достаточно высокая вероятность совершения изучаемыми моделями общих ошибок. Предложен путь минимизации их возникновения через построение оптимального запроса к модели. По результатам работы подтверждено, что на данный момент работа языковых генеративных моделей нейронных сетей требует контроля со стороны компетентного специалиста.

Ключевые слова: нейронные сети, языковая генеративная модель, ошибки нейронных сетей, цифровизация, логистика

Для цитирования: Кузьмин, А. Н. Проблема минимизации общих ошибок языковых генеративных моделей нейронных сетей в логистических системах / А. Н. Кузьмин, А. М. Туровец // Бизнес. Инновации. Экономика : сб. науч. ст. / Ин-т бизнеса БГУ. — Минск, 2024. — Вып. 10. — С. 155–160.

A. Kuzmin, A. Turovets
School of Business of BSU, Minsk, Belarus

THE PROBLEM OF MINIMIZING THE TOTAL ERRORS OF LINGUISTIC GENERATIVE MODELS OF NEURAL NETWORKS IN LOGISTIC SYSTEMS

This article considers the problem of general errors in generative language models (GLM) of neural networks. The ways of optimizing the probability of the occurrence of such errors proposed and the input and output data structure in the interaction process with those models of neural networks formulated. After the experiments conducted on the topics of probability distribution and the answers to the questions connected with the functional areas of logistics, quite a high probability of general error occurrence was discovered. Based on the results of the work, a way of optimizing the quantity of those errors was proposed. The importance of the control over generative language models is emphasized.

Keywords: neural networks, generative language model, neural network errors, digitalization, logistics

For citation: Kuzmin A., Turovets A. The problem of minimizing the total errors of linguistic generative models of neural networks in logistic systems. *Biznes. Innovatsii. Ekonomika = Business. Innovations. Economics*. Minsk, 2024, iss. 10, pp. 155–160.

Введение

Цифровизация экономических процессов способствует внедрению инноваций различного рода, что касается и логистики. Руководящий состав логистических компаний все больше обращает внимание на применение языковых генеративных, или больших языковых, моделей нейронных сетей для достижения цели минимизации уровня затрат при сохранении необходимого уровня обслуживания клиентов.

Основная часть

Нейронная сеть — математическая модель, отражающая биологические процессы, происходящие в результате мышления. Языковые генеративные модели на данный момент остаются наименее изученными. Принцип работы таких моделей опишем следующим образом: предварительное обучение осуществляется на достаточно большом массиве текстовых данных, в ходе которого сама модель запоминает наиболее часто встречающиеся устойчивые конструкции, присваивая каждому слову или группе так называемые веса или вероятности, того, что именно это слово (конструкция, выражение) будет выведено следующим после предыдущего слова. Таким образом, в общем виде разработана функциональная зависимость поведения нейронной сети от структуры запроса пользователя и особенностей получаемых выходных данных:

$$f(i, j, k) = (x_1 \cdot w_1 + x_2 \cdot w_2 + x_3 \cdot w_3 + x_n \cdot w_n) + \varepsilon \cdot w_\varepsilon,$$

где $f(i, j, k)$ — функция запроса пользователя, представляющая собой совокупность различных форматов входных данных: букв, цифр, иных типов данных); $x_n \cdot w_n$ — произведение, обозначающее вероятность выдачи той или иной смысловой конструкции (слова, словосочетания, математические символы и др.); $\varepsilon \cdot w_\varepsilon$ — случайные и общие ошибки модели в выходных данных и вероятность их возникновения.

В процессе обработки функция запроса пользователя (входные данные) преобразовываются моделью языковой генеративной нейронной сети в множество смысловых конструкций, которые генерируются с различными весами, или вероятностями, выдачи. К данному множеству добавляется некоторая ошибка ε (случайная или общая). В том случае, если ошибки нет, данное произведение равно 0, а выходные данные корректны.

Общие ошибки модели представляют собой те из них, вероятность возникновения которых пользователь может уменьшить путем создания корректного запроса с заданием соответствующих ограничений, т. е. от основных их видов — неправильной интерпретации запроса, ошибочных вычислений, лишней информации в выходных данных — можно избавиться путем своего же влияния на вероятность ее возникновения через изменение запроса [1]. Рассмотрим проблему возникновения и пути минимизации общих ошибок более подробно.

После описания структуры входных и выходных данных в рассматриваемых моделях зададим первоначальный запрос о вероятностях ошибок в поведении логистических систем любых видов к одной из моделей языковых генеративных нейронных сетей. По полученным результатам, первоначальная оценка вероятности совершения ошибок моделью находится в диапазоне 20–30 %. В оценке учтены запросы, содержащие узкоспециализированные термины, определения и понятия из различных функциональных областей логистики. Отметим, что генеративные языковые модели нейронных сетей не имеют прямого доступа к Интернету, а дают ответы только на основании тех данных, на которых они были обучены специалистами. В качестве объекта для проведения эксперимента использовалась модель Gemma 2 27B. Отметим, что какой-либо документации по принципам работы модели нет, поэтому исследования проводились через множество экспериментов с дополнительной подготовкой базы вопросов.

В рамках начального тестирования был сформирован специализированный запрос, предполагающий сбор данных на основе ответов совокупности респондентов и заданной вероятностью правильного ответа. В этом случае модель выдает распределение правильных и неправильных ответов. Например, если в качестве совокупности задать 1 000 сотрудников фокусных логистических компаний, которые отвечают на один и тот же вопрос и все они являются квалифицированными экспертами, модель выдаст случайное, но правдоподобное распределение ответов, отметив, что представленные результаты являются гипотетическими. Начальное тестирование модели необходимо для понимания реакции модели на тот или иной способ проверки ее на ошибки. В данном случае реакция языковой генеративной модели нейронной сети предсказуема, выбранный способ ответа на вопрос корректный. Это же утверждение распространяется и на задачу сбора данных об ответах респондентов на тот или иной вопрос, но с множественным выбором и без заданных вероятностей правильного ответа (модель выдаст правдоподобные результаты).

Далее для проверки была сформирована гипотеза об изменении весов параметров модели. Веса параметров влияют на выдачу той или иной конструкции после отправки пользователем запроса. Начальное предположение — веса параметров модели не изменяются в зависимости от условий и в разных сеансах взаимодействия на один и тот же набор данных будут выдавать одинаковые ответы. В качестве набора данных была задействована совокупность из 50 вопросов по логистике и ее функциональным областям [2]. Каждый правильный ответ на вопрос оценивался в один балл независимо от его типа. Частично верный ответ оценивался при сравнении с эталоном в 0,33 либо 0,5 балла. С эталоном сравнивалось два контекстных окна, формулировка запроса не изменялась. По результатам сравнения ответов с эталоном в двух сеансах взаимодействия получены следующие результаты:

- сеанс № 1 — 72 % из 100 % (с учетом частично верных);
- сеанс № 2 — 72 % из 100 % (с учетом частично верных).

Следовательно, если не изменять запрос к модели и подавать на вход в разные сеансы взаимодействия одинаковые запросы, то будут получены одинаковые ответы, т. е. содержимое сеанса № 1 равно содержимому сеанса № 2.

Можно утверждать, что веса параметров модели при сохранении всей структуры запроса не изменяются. Однако необходимо проверить еще один вариант: выдачу в том случае, если входные данные будут незначительно изменены. Результаты следующие (сравнивались сеанс № 2 и новый сеанс под № 3):

- сеанс № 2 — 72 % из 100 % (с учетом частично верных);
- сеанс № 3 — 63 % из 100 % (с учетом частично верных).

Отметим, что в 8 случаях из 50 вопросов (16 %) наблюдались расхождения в ответах между двумя сеансами взаимодействия. Следовательно, если незначительно изменить запрос, веса параметров модели (вероятности выдачи) могут изменяться в рамках одного набора данных. Результаты экспериментов показывают, что веса параметров модели до определенного момента никак не изменяются, а с некоторой точки начинают изменяться, причем значительно (отличие итогового результата между двумя сеансами в рассматриваемом случае с 50 вопросами составляет 9 %). Это отличие можно считать потерей точности выдачи.

Из вышеизложенного можно сделать вывод о том, что следует подбирать оптимальный запрос к модели за как можно меньшее количество сеансов взаимодействия во избежание возможной потери точности выдачи, как это происходит при формировании запроса к сеансу взаимодействия № 3.

Стоит учитывать, что запрос к модели на работу с набором данных из 50 вопросов был составлен без лимитирующих факторов, т. е. модель генерировала выходные данные на основе всего заложенного в нее массива. Однако наличие лимитирующих факторов в самом запросе повысило бы точность ответов модели, а в вопросе совершения ошибок следует ориентироваться на «чистый» результат без дополнительных условий.

По результатам эксперимента об изменении весов параметров модели на основе набора данных из 50 вопросов по различным функциональным областям логистики выявлено, что первоначальная оценка вероятности совершения ошибок, выданная генеративной языковой моделью нейронной сети (20–30 %) является правдоподобной именно в области ответов на вопросы по логистике и в рамках конкретной экспериментальной языковой генеративной модели, так как доля ошибок в рамках сеансов взаимодействия находится на уровне 28 % с учетом частично верных ответов.

В начальном тестировании на распределение ответов с учетом и без учета вероятностей их дачи было замечено ограничение размера контекстного окна (объема выдачи). В Gemma 2 27B оно условно равно 8 192 токенам — единицам данных для обработки входной текстовой информации. Поэтому для представления результатов ответов слишком большой совокупности модель использует последовательности. Этот момент также необходимо учитывать при работе с языковыми генеративными моделями нейронных сетей, особенно в тех случаях, когда за один раз требуется обработать большой объем информации. После заполнения контекстного окна требуется очистка сеанса или начало нового, иначе это приводит к внутренней ошибке модели.

При работе с моделью был идентифицирован новый тип ошибок, который не относится ни к общим, ни к случайным. При отправке запроса срабатывал механизм его блокировки как потенциально нежелательного. Это означает, что система внутренней фильтрации, отвечающая за этичность и безопасность использования модели, давала сбой. Новый тип ошибок можно отнести к ошибкам фильтрации. Ошибки фильтрации в своей основе содержат человеческий фактор, чем похожи на общие ошибки, но изменить частоту их возникновения может не пользователь, а специалисты, которые работают над тем, чтобы использование модели стало как можно более безопасным и соответствовало стандартам этики. Данные ошибки не настолько опасны, как случайные, так как можно незначительно изменить запрос или же лично указать на ошибку системы фильтрации через форму обратной связи.

Таким образом, в ходе эксперимента удалось выявить, что реальная вероятность совершения общей ошибки в наборе данных с вопросами, содержащими узкоспециализированные термины, находится на уровне 28 % от общего числа вопросов (случаев), а первоначальная оценка вероятности, данная моделью, – 20–30 % случаев. Максимальное отклонение прогнозной оценки от фактически полученной составляет 2 %. Отметим, что одни полные сеансы взаимодействия с моделью, т. е. первоначальный запрос, представление набора данных для дачи ответов и сами ответы, занимали от 60 % всего объема контекстного окна, иными словами, языковая генеративная модель нейронной сети работала с большим объемом информации сразу.

После определения реальной вероятности совершения общей ошибки модели на основании анализа результатов работы с составленным набором данных следует более широко рассмотреть целесообразность использования таких моделей в логистических системах. Достаточно высокая вероятность совершения общих ошибок свидетельствует о том, что на данный момент целесообразным является применение в качестве вспомогательного инструмента с большим числом направлений. Главным вопросом остается сам механизм использования. Основными путями применения моделей выступает их использование в качестве инструментов прогнозирования, с ними можно работать на операционном уровне или же контролировать качество и правильность совершенных специалистом действий в логистических процессах любой компании.

Этапы взаимодействия с моделями не отличаются в зависимости от способа (механизм сеанса взаимодействия всегда будет примерно тот же), однако на выходе будут получены совершенно разные результаты: прогноз, результат операционной деятельности специалиста и оценка качества и правильности действия соответственно. При внедрении моделей такого типа в логистические процессы важно учитывать реальную вероятность совершения ошибок, в том числе и общих, на любом из этапов работы.

Из вышеизложенного можно сделать вывод о том, что существует достаточно высокая реальная вероятность совершения моделью общей ошибки в решении узкоспециализированных задач по логистике. Тем не менее языковые генеративные модели нейронных сетей можно использовать в качестве вспомогательного инструмента по различным направлениям работы: как средство прогнозирования, облегчения операционной деятельности, инструмент оценки качества работы специалиста и т. д. Для компенсации вероятности общей ошибки потребуется специалист, занимающийся контролем выходных данных модели и при необходимости корректирующий выдачу.

С учетом найденной реальной вероятности можно рекомендовать логистическим компаниям разработку механизма повышения эффективности внедряемых в работу инновационных решений с учетом существующего на данный момент способа минимизации негативного эффекта от общих ошибок языковой модели нейронной сети. В качестве одного из предлагаемых инструментов выступает корректное структурирование запроса к любой языковой генеративной модели в рамках сеанса. Основными принципами корректного запроса выступают:

- соблюдение полноты запроса;
- наличие ограничений (лимитов);
- четкое указание формы представления данных и ожидаемых результатов.

Обеспечить полноту запроса позволяет его основная часть, содержащая запрос, без которого решение задачи моделью не имеет смысла, а также явные цели и задачи, что позволит мо-

дели понять, чего нужно достичь в рамках одного сеанса. В основную часть входит и дополнительный процесс — установка роли, которая позволяет нейросети давать ответы максимально близко к тому, как бы их давал тот или иной специалист. Например, в первой части запроса допустимо написать, чтобы модель отвечала как специалист по логистике. Иногда установка роли встречается в следующей части запроса — ограничениях, или лимитах.

Ограничения (лимиты) — элемент запроса, идущий, как правило, после основной части. Ограничения могут быть как позитивные, так и негативные: первые предполагают то, что нужно отразить в выдаче, а вторые, наоборот, указывают на то, что следует исключить из выдачи. Это позволяет модели сконцентрироваться на конкретной области массива данных, получив основную информацию оттуда, что повышает точность ответов.

Третьим этапом конструирования корректного запроса является указание формы представления данных и ожидаемых результатов. Поскольку к общим ошибкам относится, в том числе некорректное представление данных, этот момент необходимо учитывать при работе с моделями. Четкое указание объема / способа представления выходных данных и ожиданий пользователя в части результатов запроса (выходных сведений) облегчает работу самой модели, так как объем выходных данных резко снижается: например, пользователь вместо абзаца выходной информации о результатах выбора поставщика желает получить лишь наименование этого поставщика.

Соблюдение принципа полноты запроса позволит максимально исключить случаи неправильной интерпретации запроса пользователя к модели, так как будут установлены четкие цели, задачи и роли модели, что не позволит уходить в сторону в плане понимания контекста запроса. Задание корректных ограничений для модели позволит минимизировать фактические ошибки модели, так как работа с большим массивом информации в языковых генеративных моделях приводит к увеличению частоты совершения ошибок всех видов. Указание в заключительной части запроса формы представления данных и желаемых результатов повысит удобство для пользователя и потенциально снизит вероятность совершения общей ошибки, так как объем выходных данных уменьшается.

Таким образом, для снижения негативного влияния реальной вероятности совершения общих ошибок в языковых генеративных моделях нейронной сети пользователю следует придерживаться принципов постановки корректного запроса к модели, а именно: полноты запроса, наличия ограничений и указания на форму представления данных и результатов. Это позволит логистическим компаниям при внедрении и использовании данного решения в своей деятельности обеспечить максимальную его эффективность при оптимальной величине затрат на внедрение.

Заключение

В работе изучена проблема общих ошибок языковых генеративных моделей нейронных сетей и предложен путь ее решения. Выявлено, что в наборе узкоспециализированных данных, сеанс обработки которых требует от модели выделения большого количества своих ресурсов, реальная вероятность совершения общих ошибок находится на уровне в 28 % и фактически совпадает с заданной моделью прогнозной величиной в 20–30 %. Установлено, что такие модели нейронных сетей целесообразно использовать в логистике по различным направлениям, но с поддержкой специалиста, который будет компенсировать возникающие общие ошибки. Негативное воздействие минимизируется через конструирование корректного запроса к модели с соблюдением принципов полноты, наличия ограничений и указания на форму представления данных и результатов.

Список использованных источников

1. Туровец, А. М. Проблема случайной ошибки языковых генеративных моделей нейросетей при их внедрении в транспортную логистику // А. М. Туровец, А. Н. Кузьмин // Инженерное и экономическое обеспечение деятельности транспорта и машиностроения : сб. материалов VIII Междунар. науч. конф. молодых ученых, Гродно, 30 мая 2024 г. / Гродненский государственный университет им. Янки Купалы ; редкол.: Д. А. Линник (отв. ред.) [и др.]. — Гродно : ГрГУ им. Янки Купалы, 2024. — С. 667–673.

2. Тестовое задание по дисциплине «Логистика». — URL: <https://infourok.ru/testovoe-zadanie-po-discipline-logistika-3023033.html?ysclid=m1gd9v4d5p847391279> (дата обращения: 23.09.2024).

References

1. Turovets A. M., Kuzmin A. N. The problem of random error in the generative language models of neural networks in the process of implementation in transport logistics. *Inzhenernoe i ekonomicheskoe obespechenie deyatel'nosti transporta i mashinostroeniya : sbornik materialov VIII Mezhdunarodnoy nauchnoy konferencii molodyh uchenykh* [Engineering and economic support of transport and mechanical engineering : proceedings of the VIII International Scientific Conference of Young Scientists]. Grodno, 2024, pp. 667–673 (in Russian).

2. Test assignment on the logistics. Available at: <https://infourok.ru/testovoe-zadanie-po-discipline-logistika-3023033.html?ysclid=m1gd9v4d5p847391279> (accessed 23 September 2024) (in Russian).

Информация об авторе

Кузьмин Артём Николаевич — студент 4 курса, Институт бизнеса БГУ, e-mail: artem160104k@gmail.com.

Туровец Александр Михайлович — старший преподаватель кафедры логистики; Институт бизнеса БГУ, e-mail: imprudence@gmail.com.

Information about the author

Kuzmin A. — 4th year student, School of Business of BSU, e-mail: artem160104k@gmail.com.

Turovets A. — Senior lecturer at the Department of logistics; School of Business of BSU, e-mail: imprudence@gmail.com

Статья поступила в редколлегию 02.10.2024

Received by editorial board 02.10.2024