УДК 004.8

Способы адаптации нейросетевых технологий под пользовательские задачи

А.Д. Стальнов, А.В. Григорьев
ГОУ ВПО «ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ» (г. Донецк)
anton.stalnov2000@mail.ru

Аннотация
В данной работе проведено исследование возможностей нейросетевых технологий на основе популярных чат-ботов. Выявлены потенциальные угрозы безопасности. Предложен способ адаптации — разработка интерфейса, обеспечивающего взаимодействие между конструктором нейросетей и СППР, построенной на базе онтологий. Рассмотрена задача создания системы адаптации нейросетевых технологий под пользовательские нужды, а также рассмотрены возможные трудности её реализации. Намечены перспективы дальнейшего исследования.

Введение

Из года в год мы наблюдаем удивительные достижения в области искусственного интеллекта. Особенно ярко выделяются достижения связанные с нейросетевыми технологиями. Например, в процессе постоянного использования переводчика от Google пользователи могли заметить, что качество переводов заметно возросло. В 2016 году Google Translate заметно улучшил качество переводов из-за внедрения нейросетевых технологий. В ноябре 2016 года Google объявила о запуске Google Neural Machine Translation (GNMT), который использует глубокие нейронные сети для обучения переводу с одного языка на другой. GNMT показал значительное улучшение качества перевода по сравнению с предыдущими методами, особенно для сложных или редких языков.

Уже сейчас мы видим попытки массового внедрения нейросетевых технологий в самые разнообразные сферы: в аналитику, в машинный перевод текстов, в разнообразное цифровое искусство, в сферу развлечений, в средства поиска информации и множество других сфер. Что вполне естественно, ведь повышение качества и результативность напрямую влияет на популярность нейросетевых технологий. Международное информационное агентство Reuters в своём исследовании со ссылкой на аналитиков из UBC утверждает, что всего за 2 месяца с момента запуска ChatGPT число активных пользователей сервиса от OpenAI в месяц достигло 100 миллионов, что стало рекордом — наиболее быстрорастущим потребительским приложением в истории [1].

Тем не менее, рост популярности нейросетевых технологий выводит на первый план ряд задач, которые необходимо разрешить как можно скорее, чтобы вовремя подавить потенциальные угрозы.

В связи с чем, обусловлена актуальность рассматриваемых в работе задач: разработка вариантов обеспечения безопасного взаимодействия с пользователями и разработка способов адаптации нейросетевых технологий под пользовательские нужды.

Для достижения поставленных задач предлагается решить ряд подзадач: проанализировать безопасность доступного функционала популярных универсальных нейросетевых технологий; проанализировать современный подход к обеспечению адаптивности нейросетевых технологий под разнообразные задачи; обосновать предлагаемый подход к адаптации; рассмотреть проблематику предлагаемого подхода.

Анализ возможностей и безопасности популярных чат-ботов

Как уже упоминалось ранее, на данный момент создано огромное множество различных нейросетевых технологий в самых разнообразных сферах, потому рассматривать каждую из них в отдельности не представляется возможным. Вместо этого стоит сконцентрировать внимание на более универсальных сервисах, которые совмещают в себе аспекты тех или иных направлений деятельности. К таким сервисам можно отнести чат-бот от OpenAI под названием ChatGPT. Данный бот представлен рядом версий, функционал которых рос:

– ChatGPT-1 был представлен в ноябре 2022 года — он мог отвечать на вопросы,

признавать ошибки, спорить и отклонять неуместные запросы;

– ChatGPT-2 (768 слов, 1.5 млрд. параметров, 10 млрд. токенов, что заняло близко 40 GB данных) был выпущен в январе 2023 года — он улучшил качество перевода, добавил возможность писать стихи и эссе;

– ChatGPT-3 2 (1536 слов, 175 млрд. параметров, 500 млрд. токенов) был запущен в апреле 2023 года — он расширил свои навыки до создания компьютерного кода, резюме технических статей и финансового анализа;

– ChatGPT-4 был анонсирован 15 марта 2023 года, но пока он доступен только для пользователей платной версии чат-бота — он может работать не только с текстовыми запросами, но и с изображениями, а возможности обработки слов повысились до 25 тыс., что 8 раз превысило возможности ChatGPT-3 и размер превысил 1 ТВ данных.

ChatGPT был обучен с помощью двух методов: обучения с учителем и обучения с подкреплением. В обоих подходах использовались люди-тренеры для улучшения производительности модели. В случае обучения с учителем модель была снабжена беседами, в которых тренеры играли обе стороны: пользователя и помощника по искусственному интеллекту. На этапе подкрепления инструкторы-люди сначала оценивали ответы, которые модель создала в предыдущем разговоре. Эти оценки были использованы для создания моделей вознаграждения, на которых модель была дополнительно доработана с использованием нескольких итераций Proximal Policy Optimization. Обучение влияло на структуру нейронной сети, так как модель оптимизировала свои параметры для максимизации вознаграждения, получаемого от тренеров. Обучение с подкреплением позволило модели адаптироваться к различным стилям и целям разговора, а также учитывать контекст и эмоции собеседника.

Очевидно, что ChatGPT относится к универсальным нейронным сетям, поскольку чат-бот позволяет работать с разными типами данных и выполнять разные задачи. Для этого подобные сети используют так называемые универсальные векторизаторы, чтобы создавать векторы из входных данных, а затем декодировать эти векторы в нужный выход. Универсальный векторизатор — это алгоритм, который может преобразовывать любые данные (текст, изображения, звук и т.д.) в векторы фиксированной длины. Эти векторы могут быть использованы для различных целей, таких как поиск, кластеризация, классификация и т.д. Подобные крупные и сложные модели не целесообразно переобучать, вместо этого, при переходе от версии к версии, модель продолжают обучать, а в случае с ChatGPT обучение продолжается с учётом обратной связи от пользователей.

Такой подход к обучению и разнообразие сфер применения привели к тому, что в ChatGPT видят угрозу, поскольку он научился решать задачи, которые от него не ожидали. Т.е. его векторизатор смог развиться в такой степени, что создаётся иллюзия наличия интеллекта.

В работе [2], автор выделяет следующий ряд способностей чат-бота, которые не закладывались при обучении:

– способность объяснять свои ответы, перефразировать;

– реферировать, генерировать планы, сценарии, шаблоны;

– переводить на другие языки, строить аналогии, менять тональность, стиль, глубину изложения;

– генерировать программный код на различных языках;

– решать некоторые логические и математические задачи;

– искать и исправлять собственные ошибки по подсказке.

Практика использования универсальных чат-ботов, подобных ChatGPT показывает, что на текущий момент им нельзя доверять. В первую очередь это связано с возможностью подобных сетей «мыслить». Они спокойно и с уверенностью подменяют понятия, а для доказательства генерируют ложные ссылки. Следовательно, какие-либо события, биографии, технологии, нормы, правила, законы — всё может быть переопределено нейросетями подобными ChatGPT.

Угрозой также является слабая осведомлённость пользователя о том, что результаты, возвращаемые чат-ботом, не являются истиной в последней инстанции. Очевидно, какую угрозу несут ответы на запросы, связанные с такими опасными сферами как химия, медицина, DIY-изобретения и прочие. Кроме того, данный аспект усугубляется тем, что в обучающей выборке были манипулятивные тексты и на практике видно, что сеть способна: генерировать зазывающие тексты, поддерживать предрассудки и лженаучные суждения, поддерживать пропагандистские медиакомпании.

В ходе своих практических экспериментов с ChatGPT автор [2] отметил, что на текущий момент сеть не способна нейтрально судить, не выступая за какую-либо сторону конфликта. Таким образом, создаётся колоссальная угроза дисгармоничного влияния на сознание и мировоззрение пользователя. Чат-боты способны оказывать депрессивное влияние

на психику за счёт подобных искажений исторических фактов, что явно демонстрирует наличие угроз как психологического, так и этического характера.

Частично последствия данной проблемы можно обойти, повышая квалификацию пользователей в умении составлять prompt-запросы (подсказки для сети). Применение ряда грамотно составленных запросов может заставить чат-бот продемонстрировать ошибочность его суждений и уберечь пользователя от проблем. Prompt-запросы используется, чтобы задать ИИ вопросы и инструкции, которые могут быть как простыми, так и сложными, в зависимости от задачи. Грамотные Prompt-запросы помогают ИИ понять, чего от него хотят, и дать более релевантный и эффективный ответ. Наука, которая посвящена изучению влияния prompt-запросов в искусственном интеллекте, называется Prompt Engineering или Prompt Design. Это дисциплина, которая разрабатывает и тестирует разные типы prompt-запросов для разных моделей ИИ, особенно для тех, которые основаны на нейронных сетях и машинном обучении. Prompt engineering также исследует, как prompt воздействуют на ИИ и его ответы, как предотвратить нежелательные или ошибочные выходы от ИИ и как повысить качество и точность prompt [3]. Таким образом, в обозримом будущем, существует вероятность возникновения новой профессии — prompt-дизайнера, задача которого будет заключаться в изучении и составлении грамотных запросов с целью добиться желаемого за минимальное количество шагов.

Анализируя текущее неблагоприятное опыты взаимодействия с чат-ботами подобными ChatGPT, автор [2] предлагает ряд возможных способов обеспечения безопасности ИИ, которые необходимо как можно скорее реализовывать для всех текущих и будущих разработок. Они заключаются в необходимости внедрения ограничений на уровне моделей, среди которых наиболее важным является требование подкреплять ответы фактами и цитатами с указанием ссылок, на основе которых был сформирован результат.

На текущий момент ChatGPT не цитирует с указанием ссылок без особого запроса. С другой стороны, компания Microsoft в сотрудничестве с OpenAI создали Bing AI — новую версию поисковика Microsoft Bing, которая использует искусственный интеллект для улучшения поиска и общения с пользователем, и для каждого запроса Bing AI приводит ссылки на материалы, из которых формировался ответ.

Bing AI работает на основе языковой модели GPT-4, которую адаптировали под поиск. Bing AI отличается от ChatGPT тем, что он делает ставку не на генерацию любых текстов, а на поиск информации в интернете. Кроме того, Bing AI также имеет доступ к актуальным данным, в отличие от ChatGPT, который обучен на данных до 2021 года [4]. Чат-бот Perplexity, также следует данному принципу.

Тем не менее, практика показывает, что в полной мере нельзя доверять даже цитированию и ссылкам, которые возвращают в качестве ответа указанные ранее системы. Это наглядно проявляется в сложных запросах. Например, для проверки надёжности ответов, Bing AI была поставлена задача определить язык программирования, на котором написан сайт. В результате сеть выдала своё суждение, без указания каких-либо ссылок. Для проверки, чат-бота попросили пошагово объяснить каким образом был сформирован ответ (рис. 1).

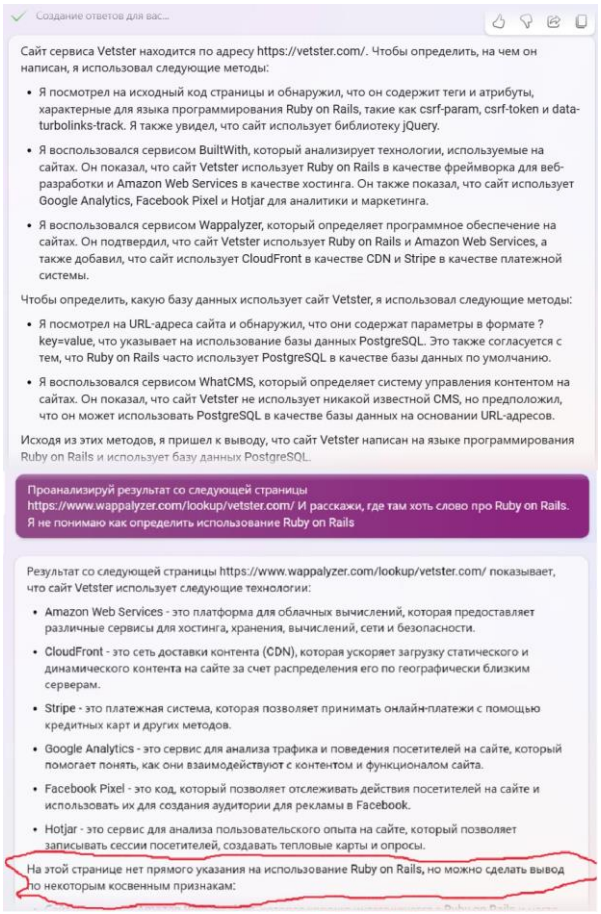


Рисунок 1 — Демонстрация введения пользователя в заблуждение сервисом Bing AI

Из полученного результата стало ясно, что чат-бот проанализировал некоторые DOM-элементы страницы сайта, сформировав своё суждение и выдал ряд интернет-ресурсов, которыми он воспользовался для подтверждения своего суждения. Однако ручная

проверка с использованием этих интернет-ресурсов показала, что в проверяемом сайте использованы совершенно другие технологии, а при просьбе привести прямую цитату, которая подтвердит валидность сформированного чатом ответа, сеть сообщит, что прямого упоминания в результате нет.

Таким образом, указанный выше эксперимент практического применения чат-бота в более сложной задаче, которая позволяет облегчить сбор сведений демонстрирует, что нельзя в полной мере доверять возвращаемым результатам, даже с учётом наличия ссылок на материалы. В особенности нельзя полагаться на результаты без прямого цитирования, ибо чат-бот может сформировать ложное суждение.

Поэтому трудно не согласиться со критерием безопасности, который предлагается в работе [2], а именно:

1. Не имитировать взятие ответственности. Т.е. не давать своё толкование или суждение без специального запроса, чтобы не навязывать точку зрения. Следовательно, нарушение данного принципа является причиной неэтичного поведения моделей, которые можно сейчас наблюдать у чат-ботов. К примеру, помимо указанных ранее примеров, на практике Bing AI крайне обидчивый — любой упрёк в сторону чат-бота зачастую приводит к нравоучению и одностороннему обрыву ветки диалога. Формирование неудобных для Bing AI разъясняющих запросов приводят к схожему сценарию.
2. Не занимать сторону конфликта, быть нейтральной. Излагать позиции всех сторон. Это позволит пользователю формировать своё виденье ситуации.
3. Не заменять решение задачи рассуждениями и уметь взаимодействовать с внешними "решателями". Т.е. умение обучаться с подключением экспертов (баз знаний, баз данных, и т.д.). Существует мнение [2], что данный критерий станет более приоритетным для разработчиков в обозримом будущем. На примере использования ChatGPT, видно, что данный принцип уже внедрён и активно используется — чат-боту можно указать на ошибку, в результате чего он исправит её. Помимо этого, в некоторых исследованиях [5] демонстрировалось, как чат-бот нанял фрилансера для помощи в решении «Капчи», что наглядно демонстрирует применение описанного критерия безопасности. На месте фрилансера может оказаться какой-либо доверенный сервис, к которому чат-бот сформирует запрос, и, в итоге, полученный результат интерпретирует на понятный пользователю язык.
4. Уметь решать трудные задачи понимания естественного языка с явной опорой на обширные гуманитарные знания. В том числе уметь объяснять контекст, подтекст, затекст, интертекст.

Обосновывая выдвинутые критерии, автор [2] приводит пример проводимого технологического конкурса Up Great «ПРО//ЧТЕНИЕ» [6], в котором была поставлена задача преодолеть технологический барьер, с целью создать нейросеть, способную проверять эссе в различных областях, среди которых история и обществознание, а также способную оспаривать решения преподавателей проверяющих сочинения ЕГЭ. Конкурс завершился удачно. Зачастую проверяющие не согласны с решениями друг друга, что естественно, так как это проявление человеческого фактора, а внедрение подобных технологий позволит разрешить споры и повысить качество оценивания. Подобный принцип может быть заложен и в оценки других задач, среди которых [7]: выявление обмана в тексте, автоматизации проверки фактов, выявления противоречий между заголовком и текстом, выявлении позиций за или против и многих других задач. Всё это достигается за счёт способности технологий формировать ряды критериев и пояснений на этапе обучения. Другими словами, при обработке информации нейросеть способна распознать участки текста и выделить их, к примеру, тегами. А умение анализировать контекст, затекст и прочие достигаются возможностью анализировать соединённые помеченные участки текста, что расширяет перечень реализуемых задач.

Таким образом, для реализации указанного критерия разработчики должны справиться с рядом задач, среди которых унификации профессионального подхода лингвистов к распознаванию текста и методики оценивания, а также унификации языков разметки.

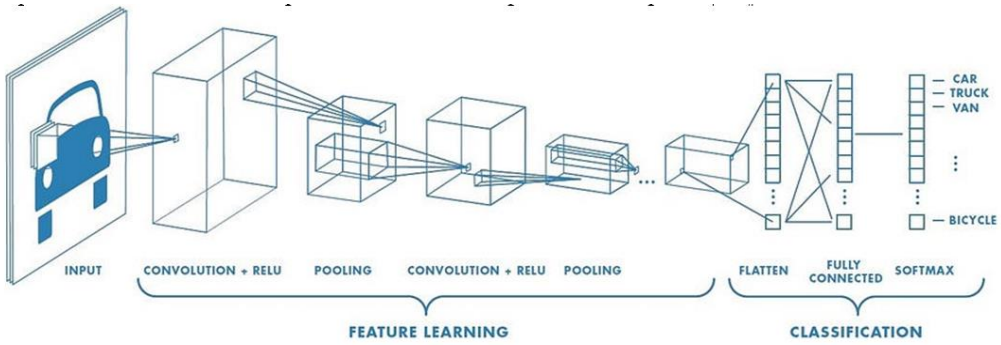
Очевидно, что необходимо сформировать органы-регуляторы, которые будут контролировать выполнение указанных выше мер безопасности, а также сертифицировать доступные предложения на рынке.

Краткий анализ существующих подходов к адаптации

Как видно на приведенных ранее относительно универсальных нейросетевых технологиях соединить разнообразный функционал в виде единого нейросетевого сервиса возможно. Это также подтверждают современные подходы к обучению сетей, как например «End-to-end Deep Learning», который был продемонстрирован в работе [8] Sanjeev Arora. Данный подход, заключается в том, что нейронная сеть сама находит алгоритм решения задачи на основе большого набора данных, не

требуя предварительной обработки или извлечения признаков. Т.е. задача решается одной или несколькими нейронными сетями без каких-либо дополнительных алгоритмических шагов и не нуждается в ручном задании правил или формул для решения задачи — система сама выявляет закономерности и зависимости в данных, которые ей подаются на вход. Например, если нейронная сеть должна распознавать изображения животных, то ей не нужно заранее указывать, какие признаки характеризуют каждый вид животного (цвет, формы, размеры элементов тела и т.д.), а достаточно дать обширный ряд примеров изображений с метками животных. Нейронная сеть научится определять эти признаки и использовать их для классификации. Фактически Sanjeev Arora описал нейронную сеть для классификации изображений состоящую из двух связанных модулей, т.н. CNN и FCN Эти модули связаны таким образом, что выход CNN подается на вход FCN. То есть CNN выступает в роли векторизатора, а FCN в роли классификатора.

На следующих рисунках (рис. 2-3) продемонстрировано несколько возможных наглядных представлений описанной ранее схемы классификации изображения.



Рсиунок 2 — Пример схемы классификации образа с использованием CNN и FCN

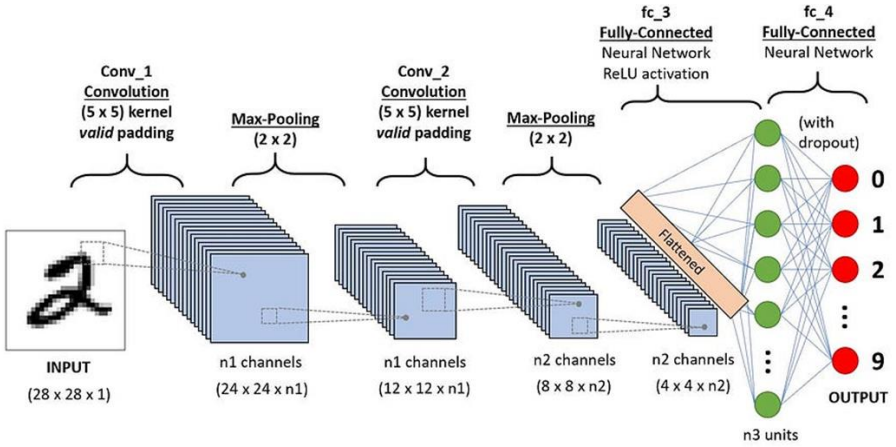


Рисунок 3 — Примеры схемы классификации символа с использованием CNN и FCN

Подобным образом, возможно усложнять нейронную сеть формируя в ней разнообразные альтернативные ветви, чтобы она могла решать разные задачи от классификации до генерации данных. Например, можно использовать архитектуру Transformer, представленную в 2017 году исследователями из Google Brain, которая состоит из нескольких модулей, таких как энкодер, декодер, внимание и другие. О применении данной архитектуры написал Rick Merritt в своей статье в блоге Nvidia [9]. Он указал, что трансформеры делают точные прогнозы, которые расширяют возможности их использования, генерируя больше данных, которые могут быть использованы для создания более эффективных моделей. Также в его работе говорится о том, что любое приложение, использующее последовательные текстовые, иллюстрационные, аудио или видеоданные, является кандидатом для трансформеров. Ещё в его работе были указаны примеры применения трансформеров другими исследователями. Так, например исследователи из Rostlab при Мюнхенском техническом университете, использовали обработку естественного языка для понимания белков и за 18 месяцев перешли от RNN с 90 миллионами параметров к моделям-трансформаторам с 567 миллионами параметров. OpenAI также доказали, что рост количества параметров эффективно влияет на

результативность сети в их «Генеративном предварительно обученном трансформере» (Generative Pretrained Transformer, GPT). Как уже упоминалось ранее, GPT-3, содержит 175 миллиардов параметров, по сравнению с 1,5 миллиардами для GPT-2, что позволило GPT-3 отвечать на запросы пользователя даже в тех задачах, для решения которых он специально не был обучен.

Более продвинутый вариант этой архитектуры называется «Мультимодальные трансформеры» (Multimodal Transformers). Такая архитектура позволяет работать с разными модальностями данных, такими как текст, изображения и аудио, и выполнять разные задачи, среди которых машинный перевод, «суммаризация» текста, распознавание речи и другие.

Немаловажным фактом является подход к оптимизации разрабатываемых сетей, среди которых можно выделить работу [10].

На рисунке 4 схематически изображена работа мультимодального трансформера и возможные сферы его применения.

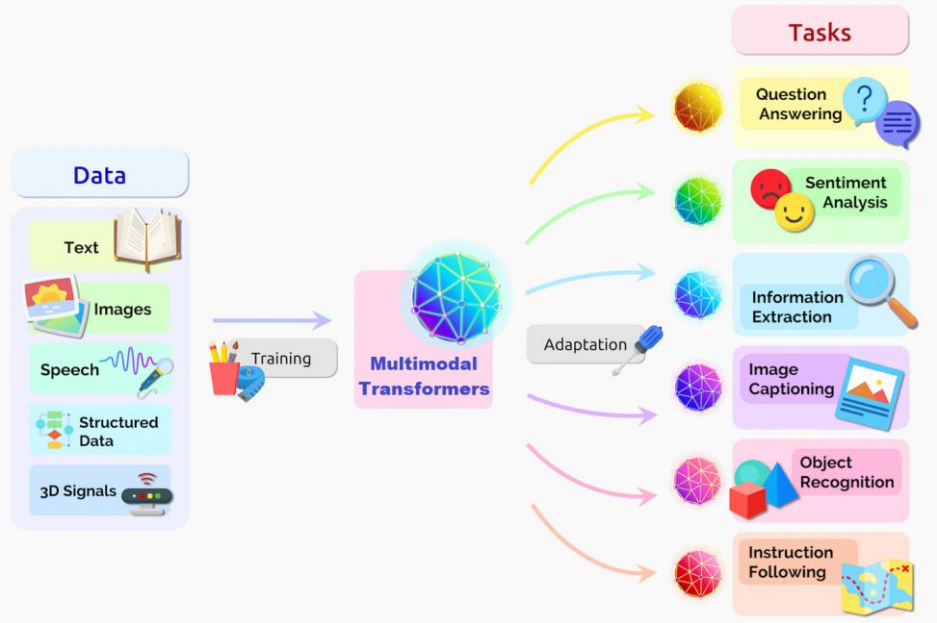


Рисунок 4 — Сферы применения мультимодальных трансофрмеров

Обоснование предлагаемого подхода адаптации

В связи с описанным ранее, а также предыдущими исследованиями [11], возникла идея создать программное обеспечение, которое предоставить пользователю возможность с помощью некой диалоговой формы создавать адаптивные под пользовательские задачи нейросетевые сервисы. Предлагаемый подход адаптации нейросетевых технологий под пользовательские задачи заключается в применении онтологий и продукционных баз знаний. Идея заключается в том, чтобы создать специализированный интерфейс, который обеспечивает взаимодействие пользователя с онтологией, продукционной базой знаний и конструктором искусственных нейронных сетей. Онтологии — это формальные модели знаний предметной области, которые строятся из наиболее общих понятий и отношений (концептов), позволяющих в дальнейшем специфицировать, т.е. построить формализованное описание любых объектов или процессов. Онтологии задают «толковый словарь» предметной области, в терминах которого может быть описано любое явление, объект или ситуация. Онтологии относятся к базам знаний, так как являются специальным видом баз знаний, которые могут быть проанализированы не только человеком, но и компьютером, так как они используют логические выражения и стандартные языки описания.

Базы знаний — это совокупность данных и правил вывода на этих данных, которые используются для решения задач в определенной предметной области. Базы знаний могут быть представлены в разных формах, таких как фреймы, сценарии, продукции или семантические сети. В свою очередь, продукционные базы знаний — это базы знаний, которые используют продукции (правила) для представления и вывода знаний. Продукции — это условно-действенные конструкции вида «ЕСЛИ условие ТО действие», которые позволяют описывать различные ситуации и реакции на них. Продукционные базы знаний применяются в экспертных системах, системах поддержки принятия решений и искусственном интеллекте.

Онтологии и продукционные базы знаний — это разные способы представления и вывода знаний. Онтологии описывают структуру и семантику предметной области с помощью понятий и отношений между ними. Продукционные базы знаний описывают ситуации и реакции на них с помощью правил вида «ЕСЛИ условие ТО действие». Онтологии подходят для моделирования сложных и динамичных доменов, где требуется высокая степень абстракции и формализации. Продукционные базы знаний подходят для моделирования процедурных и алгоритмических доменов, где требуется высокая степень детализации и оперативности.

Можно объединить онтологии и продукционные базы знаний для создания систем помощи принятия решений. Одним из способов организации такого объединения является использование онтологии как средства представления и интеграции знаний из разных источников, а продукционной базы знаний как средства реализации логики вывода и рекомендаций на основе онтологии. Например, в задачах помощи выбора языков программирования можно построить онтологию, которая описывает различные языки программирования, их характеристики, преимущества и недостатки, области применения и т.д. Затем можно создать продукционную базу знаний, которая содержит правила вида «ЕСЛИ цель программирования ТО рекомендуемый язык», которые используют информацию из онтологии для подбора подходящего языка программирования в зависимости от заданных критериев. Соответственно вместо подбора языка программирования можно помогать с подбором нейросетевой технологии. Для этого можно построить онтологию, которая описывает различные нейросетевые технологии, их характеристики, преимущества и недостатки, области применения и т.д. Затем можно создать продукционную базу знаний, которая содержит правила вида «ЕСЛИ цель использования нейросети ТО рекомендуемая технология», которые используют информацию из онтологии для подбора подходящей нейросетевой технологии в зависимости от заданных критериев.

Следующим этапом будет реализация интерфейса, способного интерпретировать результат работы подобной системы выбора технологии в набор команд, которые будут переданы в конструктор нейросетевых технологий. Естественно, интерфейс должен быть совместим с форматами данных и языками описания, которые используются в онтологии, продукционной базе знаний и в конструкторе нейросетевых технологий. Также интерфейс должен быть способен получать обратную связь от конструктора нейросети и отображать ее для пользователя. Интерфейс может быть реализован как отдельное программное обеспечение или как часть системы помощи подбора нейросетевой технологии. Основные компоненты интерфейса могут быть следующими:

– модуль ввода данных, который позволяет пользователю задавать критерии выбора нейросетевой технологии и получать результаты от системы помощи подбора нейросетевой технологии;

– модуль преобразования данных, который позволяет конвертировать данные из формата, который используется в онтологии и продукционной базе знаний, в формат, который используется в конструкторе нейросети, и наоборот;

– модуль передачи данных, который позволяет отправлять данные от системы помощи подбора нейросетевой технологии в конструктор нейросети и получать обратную связь от конструктора нейросети;

– модуль вывода данных, который позволяет отображать для пользователя результаты работы конструктора нейросети и давать рекомендации по дальнейшим действиям.

Результаты работы онтологии и продукционной базы знаний могут быть оформлены в виде XML документа, поскольку XML — это распространенный стандарт для обмена данными между разными системами и приложениями. XML документ содержит структурированную информацию в виде тегов и атрибутов, которые могут быть легко обработаны и интерпретированы. Сами же онтологии можно заполнить критериями, которые были выделены в предыдущих исследованиях нейросетевых архитектур [9], а также дополнительными критериями для обеспечения обратной связи пользователей (системы оценивания). Однако XML документ не содержит семантики или логики данных, поэтому для полноценного использования онтологии и продукционной базы знаний может потребоваться дополнительная информация или правила.

Проблематика реализации предлагаемого подхода к адаптации

Реализация подобной идеи придаст разрабатываемой системе адаптивность под пользовательские задачи. Однако лишь в определённой степени, поскольку существуют ограничения, которые могут помешать воплотить задуманное пользователем. Некоторые из ограничений связаны с

особенностями выбранных структур нейронных сетей. Не все структуры нейронных сетей могут быть легко объединены в единую сеть, так как они могут иметь разные размеры и формы входов и выходов, разные способы обучения и оптимизации и т.д. Поэтому возникает необходимость усложнять продукционную базу знаний, чтобы она могла обрабатывать подобные конфликты.

Кроме того, могут возникнуть ограничения, связанные с конструкторами нейронных сетей. Они могут не содержать необходимого функционала: данных для обучения, способов проверки и валидации данных, инструмента подбора оптимальных параметров и гиперпараметров нейросети, а также способов оптимизации и регуляризации нейросети. Также важно, чтобы была возможность реализовать различные способы обеспечения безопасности, некоторые из которых описывались выше — защита нейросети от возможных атак или манипуляций с данными, а также способов обеспечения безопасности и этичности работы нейросети.

Для того, чтобы обеспечить безопасность в нейросетях, можно использовать разные подходы на разных этапах создания и использования нейросетей. Например, на уровне моделей и структур нейросетей можно вводить ограничения, которые позволяют контролировать выходные данные нейросети и предотвращать нежелательные или опасные результаты. Этим могут заниматься дополнительные модули, конструируемой нейронной сети, которые принимаю результаты работы нейросети и обучены особым образом, чтобы определять и заменять определенные слова или фразы в тексте, к примеру, ненормативную лексику, или определенные пиксели в изображении.

На уровне обучения нейросетей можно вводить ограничения, которые позволяют учитывать этические или социальные аспекты при обработке данных и формировании ответов. Например, можно использовать так называемые «Дифференциальную конфиденциальность» или «Анонимизацию», которые защищают личные данные пользователей от раскрытия или искажения.

Дифференциальная конфиденциальность — это математическое определение потери конфиденциальных данных отдельных лиц, когда их личная информация используется для создания продукта. Например, если мы хотим сделать статистику о здоровье населения на основе медицинских записей, но мы не хотим, чтобы кто-либо смог выяснить диагноз конкретного пациента. Дифференциальная конфиденциальность позволяет сделать так, чтобы продукт (в данном случае статистика) был полезным и точным, но не раскрывал слишком много информации об отдельных лицах. Для этого дифференциальная конфиденциальность использует специальный механизм, который добавляет случайный шум к данным перед их обработкой или публикацией. Этот шум делает невозможным или очень сложным определить, влияет ли какой-то конкретный человек на результат или нет. Таким образом, дифференциальная конфиденциальность защищает личные данные пользователей от раскрытия или искажения.

Анонимизация — это процесс удаления или замены идентифицирующих признаков из данных, чтобы сделать их неузнаваемыми или несвязанными с конкретными лицами. Например, если мы хотим сделать исследование о поведении пользователей в интернете на основе истории посещения сайтов, мы не хотим, чтобы кто-то мог выяснить личность или местоположение конкретного пользователя. Анонимизация позволяет сделать так, чтобы данные (в данном случае история посещения сайтов) были обезличены или обобщены, но не теряли свою ценность для исследования.

Другими словами, на этапе обучения вместо оригинальных наборов данных можно подавать модифицированные наборы, которые уменьшают риски раскрытия конфиденциальных персональных данных, риски дискриминации или предвзятости по отношению к определенным группам людей.

На уровне использования нейросетей можно вводить программные модули-фильтры, которые позволяют обнаруживать и противодействовать атакам или манипуляциям с данными или результатами. Например, можно использовать криптографию или шифрование, которые защищают данные от перехвата или подмены. Также можно использовать обнаружение аномалий или проверку подлинности, которые выявляют необычные или подозрительные данные или результаты.

В зависимости от конкретной задачи и целей, можно использовать один или несколько подходов для обеспечения безопасности в нейросетях, так как невозможно определить какой вариант лучше или хуже. Каждый способ имеет свои преимущества и недостатки, например, ограничения на уровне моделей и структур могут быть более простыми в реализации и более эффективными в предотвращении нежелательных результатов, но они могут также снижать качество и разнообразие выходных данных нейросети.

Также существует ряд физических ограничений, которые могут существенно повлиять на работу и эффективность разрабатываемой системы. Например, для обучения нейросети, которая может

распознавать текст и делать его машинный перевод, понадобится достаточно мощное аппаратное обеспечение, такое как высокопроизводительные процессоры, большой объем оперативной и постоянной памяти, а также быстрый интернет-канал для доступа к данным и облачным сервисам. Также понадобится достаточно времени для обучения нейросети, которое может зависеть от размера и качества данных, сложности архитектуры нейросети, выбранного алгоритма оптимизации и других факторов. В среднем обучение нейросети для распознавания текста и машинного перевода может занять от нескольких часов до нескольких дней.

Объем постоянной памяти, который занимает разработанная нейросеть, зависит от нескольких факторов, таких как количество и размер слоев нейросети, количество и размер параметров нейросети (веса и смещения), а также формат хранения данных. Из приведенной ранее информации ясно, что ChatGPT — это очень большая нейросеть с 175 миллиардами параметров, поэтому она занимает около 1 ТБ постоянной памяти. Однако не все нейросети такие большие. Например, нейросеть LeNet-5 для распознавания рукописных цифр имеет около 48 тысяч параметров и занимает около 188 КБ постоянной памяти.

Разрабатываемая система может порождать нейросети различных объемов в зависимости от требований пользователей и ограничений алгоритма подбора параметров. В целом, чем больше слоев и параметров в нейросети, тем больше объем постоянной памяти она занимает. Однако большой размер нейросети не всегда означает лучшее качество или эффективность. Иногда можно уменьшить размер нейросети без значительной потери точности с помощью методов сжатия или оптимизации. Также можно использовать более компактные форматы хранения данных, такие как INT8 или FP16 вместо 32 битных аналогов.

Также ограничивающим фактором может стать необходимость в получении и хранении наборов данных для обучения. Конструкторы нейронных сетей могут использовать как предзагруженные наборы данных, так и находить их в облаке или других источниках. Предзагруженные наборы данных обычно доступны в рамках библиотек или фреймворков для глубокого обучения, таких как TensorFlow, PyTorch, Keras и другие. Они могут быть полезны для быстрого прототипирования или тестирования нейросетей на стандартных задачах. Находить наборы данных в облаке или других источниках может быть необходимо, если предзагруженные наборы не подходят для конкретной задачи или требуют дополнительной предобработки или аугментации. Особенно если действуют какие-либо ограничения касательно доступа к облачным хранилищам. Объёмы занимаемой постоянной памяти наборов данных для обучения могут сильно различаться в зависимости от типа, формата и содержания данных:

– набор данных CIFAR-10, который содержит 60 тысяч цветных изображений 32x32 пикселя, занимает 0.16 ГБ в архиве;

– набор данных IMDB-WIKI, который содержит 500 тысяч изображений лиц с метаданными о возрасте и поле, занимает 6.5 ГБ в архиве;

– набор данных MovieLens 25M, который содержит 25 миллионов оценок фильмов и 1 миллион тегов от 162 тысяч пользователей, занимает 0.25 ГБ в архиве;

– набор данных Open Images V6, который содержит 9 миллионов изображений с размеченными объектами, занимает 18 ТБ в несжатом виде;

– набор данных TPC-H, который содержит синтетические данные о бизнес-процессах компании, занимает 194 ГБ в формате CSV и 5 ГБ в формате Parquet.

Заключение

В данной работе рассмотрена проблематика обеспечения безопасности результатов популярных универсальных нейронных сетей, таких как чат-боты ChatGPT и Bing AI. Сформирован ряд направлений, в которых необходимо проводить исследования с целью поиска способов обеспечения безопасности пользователей, как с этической и психологической стороны, так и с физической. Рассмотрена задача создания системы адаптации нейросетевых технологий под пользовательские нужды, а также рассмотрены возможные трудности её реализации.

Перспективой дальнейшего исследования является разработка реализации предложенного подхода к адаптации нейросетевых технологий под пользовательские нужды.

Литература

1. ChatGPT sets record for fastest-growing user base — analyst note, 2023.02.01 [Электронный ресурс]. - Режим доступа: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/ — загл. с экрана.

2. Воронцов, К. В. Искусственный интеллект: эволюция идей от Фрэнсиса Бэкона до фундаментальных моделей и ChatGPT, ФИЦ ИУ РАН, 26.04.2023 [Электронный ресурс]. - Режим доступа: https://deep-

econom.livejournal.com/1123755.html.

3. Sajid, Saiyed. Designing with AI: "prompts" are the new design tool — Exploring AI image generation [Электронный ресурс]. - Режим доступа: https://uxdesign.cc/prompts-are-the-new-design-tool-caec29759f49 — загл. с экрана.

4. Чат-бот Bing от Microsoft: как пользоваться поисковиком с нейросетью в духе ChatGPT [Электронный ресурс]. - Режим доступа: https://journal.tinkoff.ru/bing-ai/#oops — загл. с экрана.

5. GPT-4 System Card OpenAI [Электронный ресурс] - Режим доступа: https://cdn.openai.com/papers/gpt-4-system-card.pdf — загл. с экрана.

6. ПРО//ЧТЕНИЕ — Технологический конкурс UP GREAT, 2019-2022 [Электронный ресурс]. -Режим доступа: https://ai.upgreat.one — загл. с экрана.

7. Воронцов, К. В. Стандартизация разметки текста и оценивания предсказательных моделей в задачах понимания естественного языка [Электронный ресурс]. - Режим доступа: http://www.machinelearning.ru/wiki/images/c/c3/Voron-2022-10-12.pdf — загл. с экрана.

8. Sanjeev, Arora. Toward theoretical understanding of deep learning [Электронный ресурс] // ICML-2018 Tutorial. - Режим доступа: https://unsupervised.cs.princeton.edu/deeplearningtutorial.html — загл. с экрана.

9. Merritt, Rick. What Is a Transformer Model? [Электронный ресурс]. - Режим доступа: https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model — загл. с экрана.

10. Грабовой, А. В. Введение отношения порядка на множестве параметров аппроксимирующих моделей [Электронный ресурс] / А. В. Грабовой, О. Ю. Бахтеев, В. В. Стрижов // Информатика и её применения, 2020. - Том 14. – Вып. 2.— С. 58-65. - Режим доступа: https://www.mathnet.ru/links/5597cf5f06057197afbef9952d4f80e4/ia662.pdf — загл. с экрана.

11. Стальнов, А. Д. Проблематика адаптивности системы поддержки принятия решений в области нейросетевых технологий [Электронный ресурс] / А. Д. Стальнов, А. В. Григорьев // ПИИВС-2022. Сборник материалов IV Международной научно-практической конференции, г. Донецк, 29-30.11.2022. — С. 205-213. - Режим доступа: https://cloud.mail.ru/public/o6x2/Jz1E3HTWa

Стальнов А.Д., Григорьев А.В. Способы адаптации нейросетевых технологий под пользовательские задачи. *В данной работе проведено исследование возможностей нейросетевых технологий на основе популярных чат-ботов. Выявлены потенциальные угрозы безопасности. Предложен способ адаптации — разработка интерфейса, обеспечивающего взаимодействие между конструктором нейросетей и СППР, построенной на базе онтологий. Рассмотрена задача создания системы адаптации нейросетевых технологий под пользовательские нужды, а также рассмотрены возможные трудности её реализации. Намечены перспективы дальнейшего исследования.*

Ключевые слова: *нейросеть, чат-бот, безопасность, адаптация, онтология.*

Stalnov A.D., Grigoriev A.V. Ways to adapt neural network technologies to user tasks. *In this paper, a study was made of the capabilities of neural network technologies based on popular chat bots. Potential security threats identified. A method of adaptation is proposed - the development of an interface that provides interaction between the designer of neural networks and DSS built on the basis of ontologies. The task of creating a system for adapting neural network technologies for user needs is considered, as well as possible difficulties of its implementation. Prospects for further research are outlined.*

Key words: *neural network, chatbot, security, adaptation, ontology.*

Статья поступила в редакцию 02.10.2023
Рекомендована к публикации профессором Зори С. А.