Vol. 6, No. 2, 2024

Ірина Юрчак[1], Андрій Хіч[2], Віра Оксентюк[3]

[1] Кафедра систем автоматизованого проектування, Національний університет "Львівська політехніка", вул. С. Бандери, Львів, Україна, E-mail: iryna.y.yurchak@lpnu.ua, ORCID 0009-0005-9100-8511
[2] Кафедра систем автоматизованого проектування, Національний університет "Львівська політехніка", вул. С. Бандери, Львів, Україна, E-mail: andrii.o.khich@lpnu.ua, ORCID 0009-0009-7044-3812
[3] Кафедра систем автоматизованого проектування, Національний університет "Львівська політехніка", вул. С. Бандери, Львів, Україна, E-mail: vira.m.oksentyuk@lpnu.ua, ORCID 0009-0005-1491-6946

РОЗУМІННЯ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ: МАЙБУТНЄ ШТУЧНОГО ІНТЕЛЕКТУ

Отримано: травень 20, 2024 / Переглянуто: червень 05, 2024 / Прийнято: серпень 08, 2024

https://doi.org/

Анотація. У статті проведено дослідження новітнього напрямку у штучному інтелекті - Великі Мовні Моделі, які відкривають нову еру в обробці природної мови, надаючи можливість створення більш гнучких і адаптивних систем. З їх допомогою досягається високий рівень розуміння контексту, що збагачує досвід користувачів та розширює сфери застосування штучного інтелекту. Великі мовні моделі мають величезний потенціал для переосмислення взаємодії людини з технологіями та зміни уявлення про машинне навчання. Проведено огляд історичного розвитку великих мовних моделей, зазначено компанії-лідери, що займаються науковими дослідженнями та розробкою ефективних систем. Надано інформацію щодо внутрішнього устрою та представлення знань у моделях. Висвітлено основні принципи навчання: збір даних та їх попередня обробка, вибір доцільної нейромережної архітектури, що використовується у великих мовних моделях. Зазначено, що найбільшого прогресу досягнуто з використанням нейронної мережі Трансформер, що базується на механізмі уваги. Висвітлено кроки, що значно сприяють навчанню, пост-навчанню, оптимізації швидкості навчання. Для оцінки ефективності та якості мовних моделей використовуються різні метрики, які залежать від вирішуваного завдання. Однак, незважаючи на свої переваги, великі мовні моделі на сьогодні не позбавлені проблем. Можливість генерації недостовірної інформації, вигаданих фактів та неетичних реплік представляє виклик для дослідників та розробників. Важливо продовжувати роботу над підвищенням відповідальності моделей, розробляти ефективні методи фільтрації контенту та вдосконалювати механізми навчання. Розуміння цих проблем та пошук їх рішень є ключовими кроками на шляху до створення більш ефективних та надійних великих мовних моделей. Відкритість, колективна участь та діалог між суспільством, науковою спільнотою та розробниками стають невід'ємною частиною забезпечення сталого розвитку цієї технології.

Ключові слова: великі мовні моделі, машинне навчання, глибоке навчання, набір даних, нейронна мережа Трансформер, інженерія запитів, промпт.

Вступ

У великих мовних моделей на сьогодні є кілька слабких сторін, що не дозволяють кваліфікувати ці моделі як загальний штучний інтелект (рівень можливостей мозку живих істот). Водночас нові здібності мовних моделей дозволяють дослідникам припустити, що технологічні компанії наближаються до створення сильного інтелекту швидше, ніж передбачали навіть оптимісти.

Перше покоління мовних моделей започаткувало векторне представлення слів. Ці моделі не використовувалися для подальших завдань обробки мови, тому, з точки зору обчислювальної ефективності вони є неглибокими. Прикладами таких моделей є Skip-Gram та GloVe. Отримані вектори можуть передати семантичні значення слів, але вони не залежать від контексту і не спроможні вибудовувати концепції вищого рівня.[1]

Друге покоління зосереджено на вивченні контекстних векторних представлень слів і покращенні моделей першого покоління. Прикладами таких моделей є CoVe, ELMo, OpenAI GPT та BERT. [1]

Третє покоління породило моделі, що ґрунтуються на другому поколінні, зі збільшеною продуктивністю та подоланням певних обмежень. Для моделей цього покоління можна виділити наступні характеристики: [1]

Глибоке розуміння лінгвістичних та структурних властивостей вхідних даних.

Виявлення складних семантичних відносин слів та покращене розуміння контексту.

Витягування інформації з кількох джерел: текст, зображення та аудіо.

Втілення ефективних архітектур нейромереж, застосування методів стиснення та оптимізації моделей.

Четверте покоління моделей є найновішим на сьогоднішній день. Вони збагачені наступними характеристиками:

Більший обсяг даних для навчання та більша кількість параметрів дозволяє розуміти ширший спектр мовних конструкцій та нюансів.

Покращені можливості розуміння та генерації текстів призводять до створення контекстуально точнішої відповіді.

Вдосконалення параметрів налаштування та трансферне навчання. Налаштування моделі під конкретне завдання, наприклад, переклад, реферування чи відповіді на запитання..

Розробка великих мовних моделей ведеться в різних країнах світу, причому більшість моделей надаються з відкритим вихідним кодом. На сьогоднішній день лідерами у галузі розробки та використання великих мовних моделей є:

Google використовує кілька великих мовних моделей для своєї пошукової системи.[2]

LaMDA навчена на величезному наборі даних тексту та коду, спроможна генерувати текст, перекладати мови, створювати різні види творчого контенту та відповідати на запитання інформативним чином.

PaLM 2 велика мовна модель, яка працює з чат-ботом Google Bard, і буде базовою моделлю для більшості нових функцій штучного інтелекту.

Gemini нова та потужна модель, яка може розуміти не лише текст, а й зображення, відео та аудіо. Модель здатна виконувати складні завдання з математики, фізики та інших областей, а також розуміти і генерувати високоякісний код різними мовами програмування. На даний час Gemini доступний через інтеграцію з Google Bard і Pixel 8 і поступово буде інтегрований в інші служби Google.

Google розробляє нову пошукову систему зі штучним інтелектом, яка намагається передбачити питання користувача і запропонувати більш персоналізовану відповідь. Google змінює спосіб подання результатів пошуку, додаючи чат зі штучним інтелектом, а також короткі відео та пости з соціальних мереж. Компанія відходить від традиційного списку результатів, який зробив його домінуючою пошуковою системою.

Нововведення є відповіддю на великі зміни у способах доступу людей до інформації в інтернеті, включаючи появу ботів зі штучним інтелектом. Google планує зробити свою пошукову систему більш «візуальною, зручною, особистою та людяною» з акцентом на обслуговування молодих людей у всьому світі.

OpenAI. Дослідницька компанія з розроблення комп'ютерних технологій, що фокусується на штучному інтелекті. Метою компанії є просування продуктів штучного інтелекту на користь всього людства та надання доступу до передових технологій.[3]

GPT-3 - генеративна передбачена трансформаторна модель. Модель представлено у 2020 році, з того часу швидко стала однією з найпопулярніших мовних моделей. GPT-3

використовується в широкому спектрі продуктів, включаючи чат-боти, інструменти для створення контенту та навчальні програми.

GPT-4 представлено в 2023 році. GPT-4 навчається на величезному наборі даних, що містить текст, код, зображення, аудіо, відео та інші типи даних. Ця навчальна інформація дозволяє GPT-4 розуміти зв'язки між різними типами даних та генерувати нові дані, що базуються на цих зв'язках.

OpenAI продовжує розробляти нові мовні моделі, і очікується, що вони відіграватимуть все більш важливу роль у майбутньому продуктів та програм OpenAI.

Інші великі компанії. Окрім OpenAI та Google, великі мовні моделі розробляють і інші компанії, зокрема:

Microsoft розробила мовну модель Turing NLG, яка використовується у різних продуктах Microsoft, включаючи Bing та Cortana.

Facebook розробила моделі, що використовуються у різних продуктах Facebook, включаючи Messenger та Instagram. Моделі RoBERTa, BART, XLM, BlenderBOT застосовують у соціальних мережах для модерації та рекомендацій.

IBM розробила мовну модель Watson, яка використовується в різних продуктах IBM, включаючи Watson Assistant та Watson Translate.

DeepMind, дочірня компанія Google AI розробила мовну модель Gopher у 2022 році. Модель має 280 мільярдів параметрів і є однією з найбільших мовних моделей у світі.

Ці компанії лідирують у галузі розробки великих мовних моделей з кількох причин. По-перше, вони мають доступ до великих обсягів даних, що необхідні для навчання моделей. По-друге, вони мають значні ресурси для розробки та підтримки цих моделей. По-третє, вони мають досвід у галузі машинного навчання та штучного інтелекту, які необхідні для розробки та використання великих мовних моделей.

Очікується, що конкуренція у галузі розробки великих мовних моделей продовжуватиме зростати в майбутньому. У міру розвитку технологій та збільшення обсягів даних, що використовуються для навчання моделей, очікується, що з'являться нові компанії, які зможуть конкурувати з лідерами галузі.

Огляд сучасних джерел інформації за тематикою публікації

Штучний інтелект як наука розвивається багато десятиліть поспіль. Початок розвитку пов'язують з історичним Дартмутським семінаром з питань штучного інтелекту. Семінар проводився влітку 1956 року в Дартмутському коледжі протягом 2 місяців і мав важливе значення для науки. На цей захід було запрошено молодих і амбітних вчених, що цікавилися питаннями моделювання людського розуму. Власне тоді й затверджено нову галузь науки «Artificial Intelligence»[4].

Метою семінару був розгляд питання «чи можна моделювати інтелектуальні процеси мислення і творчості за допомогою обчислювальних машин?». Як ключові питання учасники виділили: розуміння мови, самонавчання і самовдосконалення комп'ютерів. Мова допомагає оперувати знаннями та обмінюватися інформацією, тому, не дивно, що моделювання мови стало активним напрямом у багатьох дослідженнях.

Сучасний етап розвитку штучного інтелекту пов'язаний із застосуванням нейронних мереж та глибокого навчання. З революцією глибокого навчання у 2010-х почалася епоха великих мовних моделей. [5]

Великі Мовні Моделі (Large Language Models) – це моделі машинного навчання, що використовують нейронні мережі та величезні сховища даних для вирішення завдань обробки та розуміння природної мови. Вони навчаються на великих обсягах даних, застосовують мільярди параметрів, досягаючи нового рівня якості в обробці природної мови.

Великі мовні моделі вже показують визначні результати у вирішенні різноманітних завдань [6].

Завдяки аналізу великих обсягів даних, великі мовні моделі спроможні розуміти, формувати та структурувати контент в унікальному та зручному для користувачів форматі.

Великі мовні моделі дозволяють точно розуміти запити користувача природною мовою та відповідати на них. Вони аналізують контекст запиту та виконують пошук за великим масивом текстів, щоб знайти доречні відповіді на запити користувачів.

Великі мовні моделі використовують для перекладу тексту між різними мовами. У моделях використовуються алгоритми вивчення структури вхідної та вихідної мов.

Великі мовні моделі спроможні ідентифікувати та класифікувати емоційні стани та почуття у наданому тексті. Можуть виявлятися такі ознаки тексту, як позитивність, негативність, нейтральність тощо.

Можливості великих мовних моделей виходять далеко за межі того, чого їх навчали, і навіть їхні розробники не можуть зрозуміти чому. Тести показують, що системи штучного інтелекту створюють моделі реального світу подібно до людського мозку, тільки технологія у машин інша.

Багато інформації великі мовні моделі витягують з текстів, що їм надаються для глибокого навчання. Для здійснення самокорекції система шукає логіку, що лежить в основі навчальних даних, тому, чим ширший та репрезентативний набір даних, тим більше мовна модель виявляє в них загальні правила.

Термін «навчання» зазвичай є інтенсивним процесом, коли через нейронну мережу проганяють гігабайти даних і налаштовують її внутрішні зв'язки. Але глибинні сенси текстів пізнаються й під час використання великих мовних моделей. Вони вдосконалюють власні знання, використовуючи підказки користувачів. Така здатність відома як контекстне навчання, що ґрунтується на тому ж алгоритмі, як і стандартне навчання.

Один із прикладів того, як навчається велика мовна, випливає із способу, яким люди взаємодіють із чат-ботами типу ChatGPT. Моделі можна дати зрозуміти, яким чином вона має спілкуватися з користувачем і вона скорегує свої відповіді. Відповіді складаються з кількох тисяч слів, які модель бачила останніми. Як використовувати ці слова, формується фіксованими внутрішніми зв'язками моделі, але передбачається деяка варіативність.[6]

"Prompt engineering" - це методика, яка полягає у створенні ефективних та точних запитів (prompts) для мовних моделей, щоб отримувати від моделі потрібні та конкретні відповіді. Формулювання запиту може суттєво вплинути на відповіді моделі. Приклади правильно складених запитів включають зміну структури питань, додавання ключових слів, використання специфічних виразів або вказування бажаного формату відповіді. [7]

Розуміння того, як створювати ефективні запити є ключовим для використання повного потенціалу мовної моделі для керування відповідями:

Прямі запитання можуть дати конкретні відповіді. Наприклад, «Які основні характеристики великих мовних моделей?»

Запити на основі сценаріїв можна використовувати для моделювання різних процесів. Наприклад, «Я студент технічного вузу. Допоможи мені написати реферат про великі мовні моделі»

Інструкційні запити можна використовувати для виконання певних завдань. Наприклад, «Напиши 10 питань для швидкого опитування студентів на тему Великі мовні моделі»

Творчі запити можна використовувати для створення творчого вмісту. Наприклад, «Напиши коротку історію про студента, який успішно написав тест з дисципліни Великі мовні моделі».

Інший тип контекстного навчання здійснюється за допомогою створення покрокового ланцюжка відповідей. Модель просять пояснювати кожен крок своїх висновків — така тактика дозволяє успішніше вирішувати логічні та арифметичні завдання, які потребують кількох кроків. Ця процедура не була запрограмована, модель знайшла її самостійно.

Постановка проблеми

Мовні моделі навчаються на великому обсязі текстів, що збираються з різних джерел і містять різноманітну інформацію. Але така надвелика кількість несе певні проблеми. Тексти з різних джерел можуть бути не актуальними, суперечити між собою, або містити недостовірні дані. Модель споживає суперечливі дані і може генерувати неправдиву, оманливу або шкідливу інформацію.[10]

Для подолання цієї проблеми ефективними є автоматизовані методи фільтрації та перевірки фактів для усунення помилок чи невідповідностей у текстах. Для перевірки правильності, актуальності та узгодженості інформації використовувати методи аналізу знань.

Мовні моделі навчаються передбачати наступне слово або фразу в тексті на основі попереднього контексту. Це дозволяє генерувати зв'язані та правдоподібні тексти. Але, статистичне передбачення не означає логічний висновок. Модель може не перевіряти правила, факти чи докази у своїх текстах. Це може призводити до неправильної інтерпретації чи сенсу.

Вирішенням такої проблеми буде накладання на компанію-розробника відповідальності за дії та наслідки, що спричинені їх великою мовною моделлю. Варто втілювати у інтерфейс моделі певні нотатки для визначення та підтвердження джерела, авторства чи справжності текстів.

При генерації відповіді мовні моделі можуть не усвідомлювати відповідальність за свої слова. Це може призводити до генерації неетичних, маніпулятивних або злочинних текстів.

Тому, важливо закладати у функціонал моделі способи дотримання та підтримки етичних, соціальних, юридичних норм та цінностей для визначення та врахування основних принципів, правил та стандартів.

Виклад основного матеріалу

Внутрішній устрій LLM.

Великі мовні моделі є базовими моделями, які можна налаштовувати та адаптувати для вирішення широкого кола завдань. Вони мають деякі загальні характеристики: генеративні за своєю природою, використовують самостійне навчання та адаптуються до різних завдань.

Для формування результату на основі введеного запиту моделі використовують великий обсяг навчальних даних, що не мають заздалегідь проставлених міток.

Для навчання моделей використовують нейронні мережі різних архітектур та типів навчання, що дозволяє їм обробляти дані з неймовірною точністю та швидкістю. Застосовується трансферне навчання, тобто знання, що отримані з одного завдання використовують для виконання іншого завдання.

Сучасні великі мовні моделі в основному навчаються на текстових даних, але вже спроможні приймати різні вхідні дані. Це може бути необроблений текст (наприклад, повідомлення в блогах, новинні статті та повідомлення в соціальних мережах), структуровані дані (таблиці, електронні таблиці або бази даних), зображення (деякі моделі були розроблені для роботи з зображеннями шляхом перетворення зображення на текстовий опис), аудіо та відео.

Навчені великі мовні моделі можна використовувати для створення різних типів тексту: написання статей, отримання відповідей на запитання, короткого викладу книги, опис фільму або плану відпочинку на вихідні. Втім, інколи мовні моделі можуть демонструвати галюцинації. Це дивні відповіді, які не відповідають дійсності, але з часом таких відповідей меншає.

Представлення знань у моделях.

Для навчання моделі застосовуються надвеликі масиви текстів до десятків терабайтів тексту, що у певному сенсі надає універсальні знання практично про все. Знання закодовані у вагових коефіцієнтах нейронної мережі, які формуються в процесі навчання на величезних масивах текстових даних.[8]

Сама модель має величезну кількість параметрів: десятки, а іноді сотні мільярдів. Завдяки цьому можна «запам'ятати» всі стандартні конструкції великої кількості мов, включаючи мови програмування, сенс слів та термінів, стилі тексту та правила логічних висновків.

Модель навчається передбачати наступне слово в тексті на основі попередніх слів.

Модель аналізує статистичні закономірності та взаємозв'язки між словами у текстах.

Ці взаємозв'язки запам'ятовуються у вагах нейронної мережі як розподілені числові представлення слів та контексту.

Так формується "узагальнена пам'ять", що дозволяє моделі робити логічні висновки та генерувати нові формулювання на основі внутрішніх представлень мови. Знання у мовних моделях є зазвичай статистичними, а не динамічними, що виведені з наявних даних у процесі самонавчання.

Мовна модель представляє кожне слово як точку у багатовимірному складному просторі. Кожному слову чи одиниці даних зіставляється вектор фіксованої довжини (наприклад, 200-300 чисел). Ці вектори кодують семантичні приховані властивості даних. Семантично близькі слова ("кіт", "собака") матимуть схожі вектори. А у різних за змістом слів ("кіт", "комп’ютер") вектори сильно відрізняються.

Таке представлення даних у векторах називається ембедінг (embedding) і широко застосовується у NLP (Natural Language Processing) та інших завданнях машинного навчання. Ці вектори використовують для ефективної обробки даних і вони дозволяють моделі виявляти приховані закономірності.

За рахунок навчання на величезному обсязі даних в моделі близькі за змістом слова стають близькими точками, і математичні операції над ними (порівняння близькості, додавання, усереднення тощо) мають практичний зміст. Це дозволяє моделі знаходити синоніми, порівнювати зміст текстів, перефразовувати тексти. Таким чином, модель працює не з фактичними словами, а з їх сенсами.

При визначенні наступного слова його ймовірність буде залежати від сенсу всіх попередніх слів з врахуванням їхньої позиції в тексті. Цей механізм отримав назву Causal Self Attention, саме він дозволяє моделі розуміти сенс слів залежно від контексту їх використання.

Універсальна мовна модель потім часто донавчається під конкретне завдання. Завданням може бути діалог, відповіді на питання, доповнення чи редагування тексту, класифікація. Донавчання (Fine-Tuning) відбувається з використанням даних, що відображають специфіку кінцевої задачі.

Навчання мовних моделей.

Для навчання великих мовних моделей існує типовий набір кроків та методів, від підготовки даних та архітектури моделі до оптимізації та оцінки.

Формування даних для навчання. Основою успішної мовної моделі є якість та кількість навчальних даних. Різноманітний та широкий набір даних дозволяє моделі вивчати мовні нюанси, узагальнений сенс та близькість понять. Джерелами даних можуть бути книги, статті, веб-сайти, соціальні мережі та інші репозиторії з великою кількістю тексту.

Попередня обробка. Перед навчанням навчальні дані повинні бути попередньо оброблені, текст приводиться до структурованого вигляду, що забезпечує узгодженість форматів та підвищує продуктивність моделі.

Видалення зайвих символів, пунктуації, спеціальних символів, стоп-слів та інших елементів, які можуть заважати навчанню. Видалення рідкісних слів, залишаються лише інформативні слова. Видалення неінформативних фрагментів (реклама, навігація). Приведення літер до нижнього регістру для уніфікації розпізнавання та порівняння.

Токенізація - розділення тексту на дрібніші одиниці, такі як слова, частини слів або символи, яким привласнюються унікальні ідентифікатори. Модель ефективніше працює з текстом як послідовністю токенів.

Лематизація та стеммінг - це методи обробки слів, що спрямовані на приведення слів до їх базових форм. Лематизація враховує граматичні правила мови та доводить слова до словникових форм. Стемінг обрізає закінчення для зведення до однієї основи різних форм одного слова. Ці методи допомагають зменшити розмір словника та врахувати різні форми слова.

Представлення слів як числових векторів – ембендінгів (embedding), щоб модель могла працювати з даними у числовому форматі.

Структурування даних у формат, що є зручним для навчання (JSON, CSV).

Розділення даних на навчальний та тестовий набори для оцінки продуктивності моделі.

Вибір методів попередньої залежить від конкретного завдання та особливостей даних. Правильна попередня обробка текстів допомагає створити якісні та ефективні мовні моделі.

Нейромережна архітектура. Вирішальне значення для ефективного функціонування мовної моделі має вибір архітектури нейромережі. Дослідники та розробники повинні ретельно враховувати вимоги завдання, рівень складності обробки та доступні ресурси. [9]

Для побудови великих мовних моделей найчастіше використовуються наступні архітектури нейронних мереж, кожна з яких має унікальні переваги та функції:

Трансформери (Transformers) – найпопулярніша архітектура на даний момент, що заснована на механізмі уваги. Використовується у BERT, GPT-4 та інших моделях.

Рекурентні нейронні мережі (RNN) – добре працюють із послідовними даними. Використовуються у моделях типу ELMo. Нині витісняються трансформерами.

Згорткові нейронні мережі (CNN) - застосовуються для отримання локальних ознак з тексту. Використовуються як допоміжні на певних етапах обробки.

Рекурентні згорткові мережі (RCNN) – комбінують RNN та CNN, що використовувалися в деяких ранніх моделях.

Гібридні мережі – комбінація різних типів, наприклад RNN та трансформерів, що сприяє використанню переваг різних архітектур.

Найбільший прогрес у великих мовних моделях за останні роки продемонстрували трансформери. Ця архітектура демонструє найкращі результати для більшості завдань.

Навчальний процес. Навчання проводиться із використанням великого масиву високоякісних даних. Під час навчання модель ітеративно коригує значення параметрів доки модель не передбачить наступний токен з попередньої послідовності вхідних токенів. Це досягається за допомогою методів самонавчання, які вчать модель налаштовувати параметри, щоб максимально підвищити ймовірність появи наступних токенів у навчальних прикладах.

Параметри нейронної мережі, такі як кількість шарів і нейронів, прихованих шарів та механізмів уваги, мають велике значення для її ємності та продуктивності. Ці гіпер-параметри повинні бути налаштовані так, щоб забезпечити баланс між складністю та обчислювальною ефективністю, уникаючи при цьому перенавчання.

Навчені мовні моделі можна адаптувати до виконання кількох завдань із використанням наборів розмічених даних. Цей процес називається точним налаштуванням.

Розширення навчання за допомогою промптів. Застосування промптів (prompts, запити, підказки) використовується як для навчання великих мовних моделей, так і для подальшого використання у генерації тексту або вирішення інших завдань. Це стратегія, коли модель навчається з врахуванням конкретних текстових входів, які надаються їй як зразки. Для ефективного підбору промптів для роботи з конкретним завданням або контекстом використовують наступні підходи:[7]

Аналіз завдання. Вивчення характеристик та вимог конкретного завдання. Визначення типів запитань і типи потрібних відповідей.

Формулювання чітких та конкретних питань. Точне формулювання питання, щоб отримати потрібну інформацію від моделі.

Варіативність підказок. Використання різноманітних підказок, щоб отримати різні відповіді та врахувати можливі спотворення даних.

Ітеративний підхід. При використанні різних питань та підказок, аналізуються результати та коригуються підходи на основі отриманих відповідей.

Експериментування та оцінка результатів. Оцінювання відповідей моделі на основі їх відповідності очікуванням та вимогам завдання.

Важливо пам'ятати, що підбір ефективних промптів може бути ітеративним процесом, що вимагає тестування та експериментування.

Оптимізація швидкості навчання. Швидкість навчання є важливим гіпер-параметром, який контролює швидкість адаптації моделі під час навчання. Вибір потрібної швидкості навчання може значно вплинути на продуктивність моделі. Для пришвидшення процесу навчання великих мовних моделей використовують наступні методи: [6]

Попереднє навчання (Pre-Training) – модель спочатку навчається на великому корпусі даних, а потім додатково лише на цільових даних.

Квантування вагових коефіцієнтів - зменшення розрядності вагових коефіцієнтів для економії пам'яті та пришвидшення обчислень.

Розподілена обробка - паралельне навчання на кількох GPU та серверах.

Трансфер навчання - використання вагових коефіцієнтів попередньої моделі для ініціалізації вагових коефіцієнтів нової моделі.

Вибіркове навчання – оновлення лише частини вагових коефіцієнтів на кожному кроці навчання.

Збільшення розміру пакету навчальних даних - великі пакети підвищують ефективність паралельного навчання.

Підбір та комбінація цих методів дозволяє значно прискорити навчання без втрати якості моделі.

Перенавчання. Перенавчання виникає, коли модель занадто добре вивчає навчальні дані, що ставить під загрозу її здатність узагальнювати невидимі дані. Для запобігання перенавчання та

покращення можливостей узагальнення моделі використовують різні підходи, такі як зміна навчальної множини, корекція вагових коефіцієнтів та примусова зупинка навчання.

Оцінка продуктивності моделі. Для оцінки ефективності та якості мовних моделей використовуються різні метрики:[1]

Точність (Accuracy) – частка правильних відповідей моделі на тестовому наборі даних. Застосовується для завдань класифікації.

F-мера (F1 Score) - усереднена міра точності та повноти для завдань класифікації.

Перплексія (Perplexity) – показник, наскільки добре розподіл ймовірностей або ймовірнісна модель передбачає зразок. Низька перплексія свідчить, що розподіл ймовірностей добре передбачає вибірку.

BLEU (Bilingual Evaluation Understudy) – це вимір відмінностей між автоматичним перекладом та еталонним перекладом текстом, що виконано людиною. Чим вище значення показника, тим краще.

ROUGE (Recall-Oriented Understudy) – це набір показників, що використовується для оцінки автоматичних програм автореферування та машинного перекладу при обробці природної мови.

CIDEr (Consensus-based Image Description Evaluation) - це важливий інструмент для тестування та покращення алгоритмів генерації текстових описів у завданнях комп'ютерного зору та обробки природної мови. Оцінює схожість сенсу згенерованого та еталонного тексту.

Час виконання – наскільки швидко модель обробляє запити.

Обсяг обчислювальних ресурсів - кількість дискової та оперативної пам'яті, обчислювальних ядер тощо.

Вибір конкретної метрики залежить від вирішуваного завдання і важливих аспектів якості для неї.

Навчання великих мовних моделей потребує ретельної уваги до деталей та глибокого розуміння основних методів. Це складний процес, який включає відбір та попередню обробку даних, вибір відповідної архітектури моделі, оптимізацію процесу навчання та оцінку продуктивності за допомогою відповідних метрик та контрольних показників. Дослідники та розробники постійно прагнуть покращити та розширити можливості великих мовних моделей.

Важливість ефективних методів навчання лише зростатиме, відкриваючи справжній потенціал мовних моделей та призводячи до нових додатків та рішень, які перетворять область обробки природної мови

Висновки

Великі мовні моделі відкривають нову еру в обробці природної мови, надаючи можливість створення більш гнучких і адаптивних систем. З їх допомогою досягається високий рівень розуміння контексту, що збагачує досвід користувачів та розширює сфери застосування штучного інтелекту. Великі мовні моделі мають величезний потенціал для переосмислення взаємодії людини з технологіями та зміни уявлення про машинне навчання.

Вони можуть бути корисні для різних цілей та програм, пов'язаних з мовою. Однак, вони також є складним і загадковим об'єктом для вивчення і розуміння. Важко зрозуміти, звідки вони знають, що вони розуміють у мові і як вони взаємодіють з людьми.

Можливість генерації недостовірної інформації, вигаданих фактів та неетичних реплік представляє виклик для дослідників та розробників. Ці проблеми та виклики вимагають подальшого дослідження та розробки методів та рішень, які допоможуть краще зрозуміти, контролювати та використовувати великі мовні моделі. Важливо продовжувати роботу над підвищенням відповідальності розробників моделей, розробляти ефективні методи фільтрації контенту та вдосконалювати механізми навчання.

Розуміння цих проблем та пошук їх рішень є ключовими кроками на шляху до створення більш ефективних та надійних великих мовних моделей. Відкритість, колективна участь та діалог між суспільством, науковою спільнотою та розробниками стають невід'ємною частиною забезпечення сталого розвитку цієї технології.

Великі мовні моделі обіцяють стати наріжним каменем майбутнього штучного інтелекту. Розробники, вчені та суспільство загалом мають можливість спільно управляти цим розвитком, забезпечуючи максимальну вигоду та мінімізуючи потенційні ризики. Попереду на людство чекають захоплюючі виклики і можливості в світі штучного інтелекту, що стрімко змінюється.

Перелік використаних джерел

[1] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu. A Survey on Evaluation of Large Language Models [Online] URL: https://dl.acm.org/doi/pdf/10.1145/3641289 (Accessed: 02/05/2024).

[2] Large Language Models powered by world-class Google AI [Online] URL: https://cloud.google.com/ai/llms

[3] OpenAI Large Language Models [Online] URL: https://platform.openai.com/docs/models/ (Accessed: 02/05/2024).

[4] AI history: the Dartmouth Conference. [Online] URL: https://www.klondike.ai/en/ai-history-the-dartmouth-conference/, (Accessed: 02/05/2024).

[5] A Very Gentle Introduction to Large Language Models without the Hype [Online] URL: https://mark-riedl.medium.com/a-very-gentle-introduction-to-large-language-models-without-the-hype-5f67941fa59e, (Accessed: 02/05/2024).

[6] Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann. ChatGPT for good? On opportunities and challenges of large language models for education, Learning and Individual Differences, Volume 103, 2023, 102274, ISSN 1041-6080, https://doi.org/10.1016/j.lindif.2023.102274.

[7] Jiaqi Wang, Zhengliang Liu, Lin Zhao, Review of large vision models and visual prompt engineering, Meta-Radiology, Volume 1, Issue 3, 2023, 100047, ISSN 2950-1628, https://doi.org/10.1016/j.metrad.2023.100047.

[8] Usman Naseem, Imran Razzak, Shah Khalid Khan, Mukesh Prasad. A Comprehensive Survey on Word Representation Models: From Classical to State-of-the-Art Word Representation Language Models. ACM Transactions on Asian and Low-Resource Language Information ProcessingVolume 20Issue 5Article No.:74 pp.1–35 https://doi.org/10.1145/3434237

[9] Jakob Uszkoreit. Transformer: A Novel Neural Network Architecture for Language Understanding. [Online] URL: https://blog.research.google/2017/08/transformer-novel-neural-network.html (Accessed: 02/05/2024).

[10] Tamkin, A., Brundage, M., Clark, J., & Ganguli, D. (2021). Understanding the capabilities, limitations, and societal impact of large language models. arXiv preprint arXiv:2102.02503. https://doi.org/10.48550/arXiv.2102.02503.

Iryna Yurchak[1], Andrii Khich[2], Vira Oksentyuk[3]

[1] Computer Design Systems Department, Lviv Polytechnic National University, Ukraine, Lviv, S. Bandery street 12, E-mail: iryna.y.yurchak@lpnu.ua, ORCID 0009-0005-9100-8511

[2] Computer Design Systems Department, Lviv Polytechnic National University, Ukraine, Lviv, S. Bandery street 12, E-mail: andrii.o.khich@lpnu.ua, ORCID /0009-0009-7044-3812

[3]Computer Design Systems Department, Lviv Polytechnic National University, Ukraine, Lviv, S. Bandery street 12, E-mail: vira.m.oksentyuk@lpnu.ua, ORCID 0009-0005-1491-6946

UNDERSTANDING LARGE LANGUAGE MODELS: THE FUTURE OF ARTIFICIAL INTELLIGENCE

Recieved: May 20, 2024 / Revised: June 05, 2024 / Accepted: August 08, 2024

Abstract. The article examines the newest direction in artificial intelligence - Large Language Models, which open a new era in natural language processing, providing the opportunity to create more flexible and adaptive systems. With their help, a high level of understanding of the context is achieved, which enriches the user experience and expands the fields of application of artificial intelligence. Large language models have enormous potential to redefine human interaction with technology and change the way we think about machine learning. An overview of the historical development of large language models is carried out, leading companies engaged in scientific research and development of effective systems are indicated. Information is provided regarding the internal structure and representation of knowledge in models. The main principles of

learning are highlighted: data collection and their pre-processing, selection of an appropriate neural network architecture used in large language models. It is noted that the greatest progress has been achieved using the Transformer neural network, which is based on the mechanism of attention. The steps that significantly contribute to training, post-training, and optimizing the speed of training are highlighted. To evaluate the effectiveness and quality of language models, various metrics are used, which depend on the task to be solved. However, despite their advantages, large language models today are not without problems. The possibility of generating false information, fabricated facts, and unethical remarks presents a challenge for researchers and developers. It is important to continue work on increasing the responsibility of models, develop effective content filtering methods, and improve learning mechanisms. Understanding these problems and finding solutions to them are key steps towards building more efficient and reliable large language models. Openness, collective participation and dialogue between society, the scientific community and developers are becoming an integral part of ensuring the sustainable development of this technology.

Keywords: Large Language Models, Machine Learning, Deep Learning, data set, Transformer Neural Network, Prompt Engineering.