УДК 004.048

Я.А. Кулик
В.В. Войцеховський

АНАЛІЗ МЕТОДІВ ТА ПІДХОДІВ У ВИКОРИСТАННІ ТЕХНОЛОГІЙ ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ ОБРОБКИ ТЕКСТОВОЇ ІНФОРМАЦІЇ НА ПРИКЛАДІ МОВНОЇ МОДЕЛІ PaLM

Вінницький національний технічний університет

Анотація

У цій статті розглядаються методи та підходи, які використовуються для обробки текстової інформації з використанням технологій штучного інтелекту на прикладі впровадження мовної моделі PaLM. Проаналізовані перспективи розвитку технології PaLM, висвітлені сильні та слабкі сторони використаних підходів у завданнях обробки тексту та в галузі NLP.

Ключові слова:

Штучний інтелект; обробка природної мови; мовна модель PaLM; аналіз тексту; машинне навчання; глибинне навчання; технології штучного інтелекту.

Abstract

This article examines the methods and approaches used for processing textual information using artificial intelligence technologies on the example of the implementation of the PaLM language model. The prospects for the development of PaLM technology are analyzed, the strengths and weaknesses of the approaches used in text processing tasks and in the field of NLP are highlighted.

Keywords:

Artificial Intelligence; natural language processing; PaLM language model; text analysis; machine learning; deep learning; artificial intelligence technologies.

Вступ

У сучасному машинному навчанні виникає збільшена потреба в аналізі та обробці текстової інформації. Правильне озуміння тексту має значний вплив на обробку вхідних даних та формування результатів. Трансляція сенсу мовлення становить складне завдання через різноманітність особливостей людської мови. Ці проблеми вирішуються у сфері обробки природної мови (NLP), що на сьогодні вже має широкий арсенал інструментів.

Мовна модель Pathways Language Model від компанії Google є одним із значущих інструментів в NLP. PaLM володіє багатьма функціями, серед яких – розуміння загальних логічних тверджень, арифметичне мислення, роз’яснення жартів, створення коду та переклад. У поєднанні з ланцюжком мислення, PaLM показала істотно кращі результати на наборах даних, які потребують розуміння на кількох рівнях, ніж конкурентні моделі. PaLM відіграє ключову роль у розвитку новітніх методів аналізу тексту та у покращенні результатів, отриманих у галузі NLP.

Принципи роботи моделі PaLM

Модель PaLM використовує мережу, яка базується на нейромережевому підході, відомому як трансформер. У загальних рисах PaLM схожа на конкуруючі моделі на основі трансформера, включаючи моделі OpenAI, такі як GPT-3 та GPT-4. Мережа трансформера – це новаторська архітектура, яка спрямована на вирішення послідовних завдань, одночасно з легкістю обробляючи залежності з широким спектром зв'язків [1].

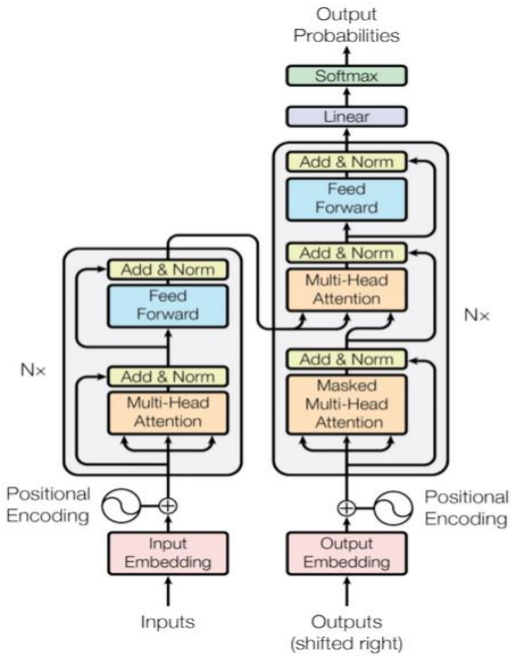

Рис 1 – Алгоритм роботи мережі-трансформера

На рис 1 зображено як мережа-трансформер може обробляти всі слова речення та визначати сенс слів одночасно. Для початку ця мережа перетворює слова у вектори подібно до словника, де слова схожих значень групуються разом. Кожне слово, відповідно до його значення, відображається і отримує певне значення, таким чином формуючи вектор. Дана модель вирішує одну з найголовніших проблем – в різних реченнях кожне слово може мати різні значення. Тому для вирішення цієї проблеми використовуються позиційні кодери. Це вектори, які надають контекст залежно від положення слова в реченні. Фактично можна описати цей процес як:

Слово → Сенс → Позиційне вбудування → Кінцевий вектор, уявлений як Контекст.

Для кожного слова створюється вектор уваги, який зафіксовує контекстуальний зв'язок між словами у цьому реченні. Визначається кілька векторів уваги для кожного слова та береться зважене середнє для обчислення кінцевого вектора кожного слова (рис 2).

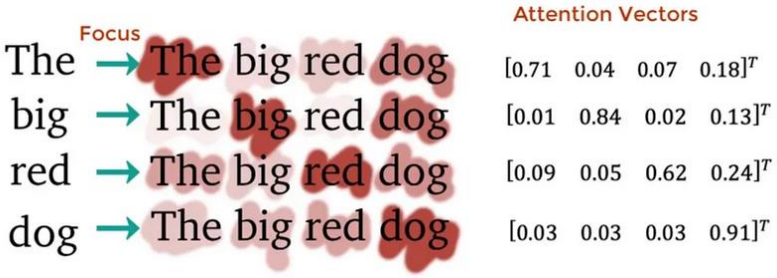

Рис 2 – Процес формування вектора уваги

Наступний крок – це нейромережа прямого поширення. Просту нейромережу прямого поширення застосовують до кожного вектора уваги, щоб перетворити їх у форму, яка прийнятна для наступного рівня кодера або декодера. Завдяки цьому можливо передати всі слова одночасно до блоку кодування й отримати набір закодованих векторів для кожного слова одночасно – наприклад: під час тренування перекладу з англійської мови на українську потрібно надати речення англійською разом із його перекладеною українською версією для того, щоб модель могла вчитися. Таким чином, речення англійською проходять через блок кодування, а українською – через блок декодування.

Кожен вектор передається у вхід до блоку зворотного розповсюдження (feed-forward unit), це перетворює вихідні вектори у форму, яка легко прийнятна іншим блоком декодування або лінійним шаром. Лінійний шар – це ще один блок прямого поширення, який розширює розмірність до кількості слів в українській мові після перекладу. Декодер працює аналогічно, але генерує одне слово за раз, зліва направо. Він уважно спостерігає не тільки за раніше створеними словами, але й за кінцевими представленнями, що створені кодером [2].

Моделі, що використовують такий підхід широко використовуються у сучасному світі – серед них GPT-3, GPT-4, BLOOM, BERT, ViT та інші. Зазвичай такі моделі вирішують різноманітні завдання, такі як генерація тексту, розуміння мови, переклад, генерація коду та аналіз вмісту. Модель PaLM відзначається у своїй здатності до багатьох функцій завдяки уніфікації трьох ключових технологічних підходів.

1. PaLM використовує техніку обчислювально-оптимального масштабування, яка дозволяє масштабувати розмір моделі та навчальний набір даних таким чином, щоб отримати більш ефективні результати, забезпечуючи швидший інференс (використання навченої моделі для отримання вихідних результатів на нових, раніше не бачених даних) та зменшення обсягу обслуговування моделі.
2. PaLM має вдосконалений попередній набір даних, включивши більше мовних ресурсів, таких як сотні людських та програмних мов, математичні рівняння, наукові статті та веб-сторінки. Це дозволяє моделі краще адаптуватися до різних завдань та виконувати більш широкий спектр функцій.
3. PaLM має покращену архітектуру, оскільки була навчена на найсучасніших різноманітних задачах, що дозволило моделі краще зрозуміти різні аспекти мови та різноманіття завдань, що потребуватимуть вирішення [3].

Вихідний набір даних для початкової тренувальної моделі PaLM включав 780 мільярдів токенів, які охоплювали тексти з різних джерел: соціальні медіа (50%), веб-сайти (27%), новинні статті (1%), Вікіпедія (4%) та вихідний код (5%). Вихідний код відбирався за ліцензіями, які обмежують відтворення коду, що має ліцензію GPL. Розподіл тематики текстів схематично зображено на рис 3.

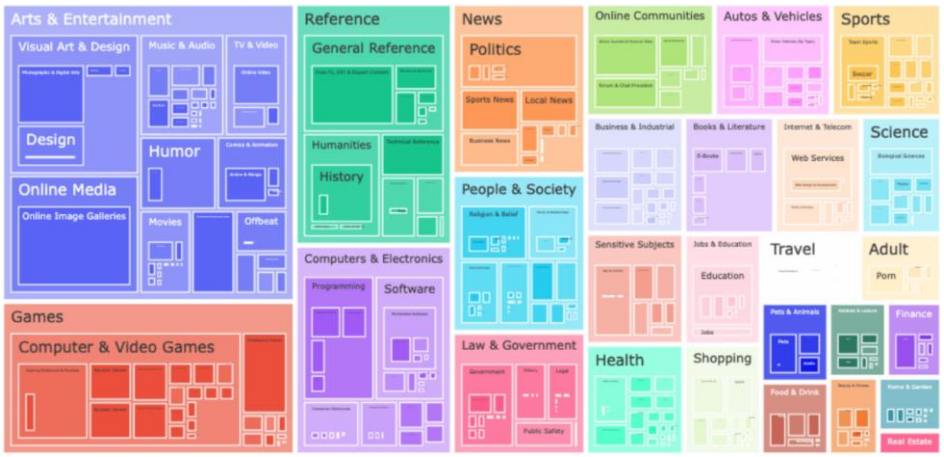

Рис 3 – Розподіл тематики текстів для моделі PaLM

Приклади застосування моделі PaLM

PaLM – і, зокрема, PaLM 2 – може виконувати багато функцій, включаючи наступні:

1. Генерація тексту. Створює текст на будь-яку тему, використовуючи текстовий запит.
2. Узагальнює великі обсяги контенту до більш компактної форми.
3. Аналіз контенту. Ця функція допомагає користувачам розуміти, що міститься в блоці контенту. Це може включати аналіз настроїв, щоб визначити, чи є тон контенту позитивним чи негативним.
4. Мислення. У PaLM є різноманітний набір даних, який охоплює наукові статті та контент з математичними виразами. Цей набір даних покращує вміння моделі в логіці та мисленні здорового глузду щодо наборів проблем, що надаються через запит.
5. Генерація коду. PaLM генерує програмний код на 80 різних мовах програмування, включаючи популярні мови такі як: Java, JavaScript та Python.
6. Аналіз коду. Модель може розглянути блок коду та ідентифікувати потенційні помилки в коді.
7. Переклад тексту. PaLM може виконувати переклад тексту.

Модель PaLM має потенціал застосування в різних галузях, включаючи:

1. Генерація тексту: PaLM може бути використаний для створення тексту на різні теми за запитом. Це відкриває можливості для створення контенту для блогів, статей, веб-сайтів та іншого.
2. Узагальнення: Модель здатна стисло узагальнювати обсяги великих текстових даних, що може бути корисним для підготовки керівницьких звітів, рефератів тощо.
3. Аналіз відгуків користувачів: PaLM може допомогти у розумінні контексту та настроїв у відгуках користувачів. Він може виявляти позитивні, негативні або нейтральні настрої та допомагати у відсіюванні важливих даних з великого потоку відгуків. Модель може розпізнавати настрої, оцінювати загальний тон відгуку та виокремлювати ключові аспекти, які вказують на задоволеність або незадоволеність користувача. Це може бути корисним для підприємств для удосконалення своїх товарів або послуг на основі зібраних даних з відгуків користувачів [4].

Крім того, також було досліджено нові можливості та майбутні перспективи PaLM на бенчмарку "Beyond the Imitation Game" (BIG-bench), недавно випущеному наборі з більш ніж 150 нових завдань

для мовленнєвих моделей, і було встановлено, що PaLM досягає значно вищих результатів, в порівнянні її конкурентів на ринку. Було порівняно продуктивність PaLM з Gopher та Chinchilla на загальному наборі з 58 таких завдань. Важливо відзначити, що продуктивність PaLM залежить від масштабу та відповідає логарифмічному закону, схожому на попередні моделі, що свідчить про те, що покращення результатів від масштабу ще не досягли плато. При цьому 5-shot версія PaLM 540B продемонструвала кращі результати, ніж середній показник продуктивності людей, які намагалися вирішити ті ж завдання. Результати проілюстровано на рис. 4.

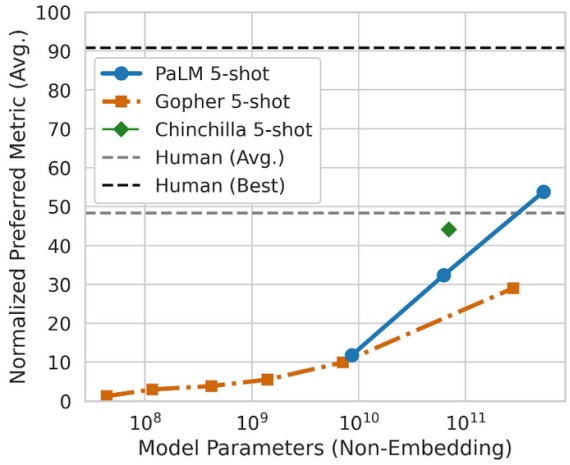

Рисунок 4 – Поведінка масштабування PaLM на наборі завдань BIG-bench.

Переваги та обмеження досліджуваної моделі

На базі проведеного дослідження у попередніх розділах можна виділити наступні переваги моделі PaLM:

1. Ефективне навчання: PaLM відзначається здатністю паралельно оптимізувати процес навчання. Це призводить до швидшого тренування моделі та зменшення вимог до обчислювальних ресурсів порівняно з традиційними мовними моделями.
2. Покращене розуміння мови: використовуючи паралельну, PaLM істотно поліпшує розуміння контексту, що дозволяє генерувати більш відповідні відповіді в різних завданнях обробки природної мови.
3. Масштабованість: моделі PaLM спроектовані для ефективного масштабування, роблячи їх відповідними для різноманітних застосувань – від розробки чат-ботів до складної аналітики даних.

Зважаючи на обмеження та особливості моделі PaLM, слід відзначити наступне [5]:

1. Використання: PaLM є моделлю, розробленою Google і опублікованою компанією. Хоча з виходом PaLM 2 Google відкрив часткові можливості для зовнішніх розробників через API, Firebase та на Colab, комерційні умови використання поки не є зовсім визначеними. Зовнішні розробники не можуть вносити новий код або брати участь у розвитку PaLM через його закритий характер і відсутність відкритості.
2. Зображення: PaLM 2 може демонструвати візуальні результати в рамках запиту, але він не здатний повністю генерувати нові зображення самостійно. Зокрема, хоча інструменти, побудовані на базі PaLM 2 (наприклад, Bard), можуть підтримувати розширення з

підтримкою інших сервісів, наприклад, з Adobe Firefly, це потребує підтримки розробників.

3. Обґрунтованість: PaLM є закритою моделлю і не надає достатньо деталей для визначення пояснень, що є важливим для розуміння користувачами та організаціями того, як модель прийшла до конкретного рішення. Це важливо для забезпечення довіри до моделі та її результатів.
4. Токсичний контент: Однією з ключових проблем PaLM є ризик наявності токсичного контенту, який може містити упередженість, зловмисність або шкідливість для користувачів. Це може стати важливим фактором ризику при використанні моделі для аналізу або створення контенту.

Це перелік обмежень та питань, які потребують уваги при використанні моделі PaLM.

Висновок

Під час даної роботи було докладно проаналізовано модель PaLM у всій її глибині та різноманітності. Описано не лише базові принципи її функціонування, а й конкретні сценарії застосування в різноманітних галузях, включаючи генерацію тексту, переклад, аналіз відгуків користувачів, мовні можливості та багато інших. Підкреслено ключові переваги цієї моделі, такі як ефективність у навчанні, вдосконалене розуміння контексту та здатність генерувати більш контекстуально-значущі відповіді у завданнях обробки природної мови. Водночас були визначені деякі недоліки, такі як обмеження в поясненні прийняття рішень та потенційні проблеми з токсичним вмістом.

Аналізуючи перспективи моделі PaLM, було виявлено широкий потенціал застосування в різних галузях та підтверджено можливості її подальшого вдосконалення та розвитку для вирішення викликів у сфері обробки природної мови.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. S. Narang and A. Chowdhery (2022). — Pathways Language Model (PaLM). Режим доступу: https://blog.research.google/2022/04/pathways-language-model-palm-scaling-to.html

2. U. Ankit (2022)— Transformer Neural Networks: A Step-by-Step Breakdown. Режим доступу: https://builtin.com/artificial-intelligence/transformer-neural-network

3. Z. Ghahramani (2023) — Introducing PaLM 2. Режим доступу: https://blog.google/technology/ai/google-palm-2-ai-large-language-model/

4. A. Vaswani, N. Shazeerm, N. Parmar, J. Uszkoreit, L. Jones, Aidan N. Gomez, Ł. Kaiser, I. Polosukhin (2017) — Attention Is All You Need. Режим доступу: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

5. H .Hapke (2023) — Introduction to Google's PaLM 2 API. Режим доступу: https://digits.com/developer/posts/introduction-to-googles-palm-2-api/

Науковий керівник – Кулик Ярослав Анатолійович, доцент кафедри автоматизації та інтелектуальних інформаційних технологій, Факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, м.Вінниця, e-mail: kulyk.y.a@vntu.edu.ua

Войцеховський Вільям Вільямович – студент групи 1ІСТ-22м, кафедра автоматизації та інтелектуальних інформаційних технологій, Факультет інтелектуальних інформаційних технологій

та автоматизації, Вінницький національний технічний університет, м.Вінниця, e-mail: fkca.1akit18.VVV@gmail.com

Academic supervisor – Yaroslav Anatoliyovych Kulyk, Associate Professor of the Department of Automation and Intelligent Information Technologies, Faculty of Intellectual Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail: kulyk.y.a@vntu.edu.ua

Voitsekhovskyi Viliam Viliyamovych – student of 1IST-22m group, Department of Automatization and Intelligent Information Technologies, Faculty of Intellectual Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail: fkca.1akit18.VVV@gmail.com