СЕКЦІЯ 2. ІНТЕЛЕКТУАЛЬНІ ІНФОРМАЦІЙНІ СИСТЕМИ ТА МАШИННЕ НАВЧАННЯ

ОБРОБКА ПРИРОДНОЇ МОВИ ЗА ДОПОМОГОЮ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ / NATURAL LANGUAGE PROCESSING WITH BIG LANGUAGE MODELS

Задохін Д.В., Глибовець М.М. / Zadokhin D.V., Glybovets M.M.
Національний університет «Києво-Могилянська академія» / National University of Kyiv-Mohyla Academy
09117, м. Біла Церква, вул Вокзальна 11, кв. 157. Тел +380970809841
Email: d.zadokhin@ukma.edu.ua

This report was created to overview the usage of large language models for resolving natural language processing tasks. In the course of this paper, we will review mechanisms which allow models with Transformer architecture to be effective for NLP task. We will highlight how Transformer structure made possible of Large Language Models and look at the examples of such models. At the end of the paper, we will review the structure of open-source model Meta Llama and its usage for text generation with limited computational resources.

Обробка природньої мови (NLP) - це один з напрямків досліджень у сфері штучного інтелекту, що полягає у створення інструментів для обробки, аналізу і генерації людської мови. Вирішення проблем NLP можу потенційно стерти бар'єр у взаємодії між людиною і комп'ютером, адже для людей основний засіб комунікації - це мова, в той час як комп'ютерні системи в основному сприймають інформацію за строгими протоколами.

Алгоритми машинного навчання широко застосовуються до вирішення задач обробки природної мови. До 2017 року рекурентні і згорткові нейронні мережі вважались найефективнішими для таких задач як машинний переклад і моделювання мови. В 2017 з'явилась нова архітектура глибоких нейронних мереж - Transformer, що виявилась надзвичайно ефективною для вирішення задач NLP [1].

Оригінальна модель Transfomer складається із двох основних компонентів: енкодера і декодера. Енкодер перетворює вхідну символьну послідовність слів на послідовність векторних представлень такої ж довжини. Декодер - з послідовності, згенерованої енкодером для вхідної послідовності токенів, послідовно генерувати послідовність результуючих токенів. Сучасні моделі на основі Transformer можуть мати тільки енкодер чи тільки декодер, в залежності від вирішуваних задач.

В основі архітектури Transformer лежить механізм уваги. Він встановлює зв'язки між різними словами однієї послідовності і використовує ці звязки для обчислення їх представлень. Самоувага (self-attention) успішно застосовувалася в різних задачах, включаючи розуміння прочитаного, абстрактне узагальнення, текстові висновки та навчання незалежних від задач представлень речень і до появи Transformer. Особливість архітектури полягає у тому, що вона цілком покладається на цей механізм, не використовуючи рекурентну структуру чи згорткові фільтри.

Multi-Head Attention (досл. багатоголова увага) - підхід до уваги, вперше використаний саме в архітектурі Transformer. Ідея полягає в тому, щоб порівнювати між собою не повнорозмірні вектори (розмірність моделі d_{model}), а їхні проекції на простори менших розмірностей (d_k). При цьому, розмірність цих просторів повинна бути кратно менша від розмірності векторних представлень токенів (ембедингів): $d_{model} = d_k \cdot h$. Параметр h - це кількість центрів уваги.

Тип шару	Обчислювальна складність	Послідовні операції
Self-Attention	$O(n^2 \cdot d)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(\log_k(n))$

Таблиця 1. Порівняльна характеристика різних типів шарів у глибоких нейронних мережах. n – кількість слів у послідовності, d - розмірність векторних представлень токенів, k - розмір згорткового фільтру у згортковому шарі.

Із таблиці 1 видно, що складність обчислень на Self-Attention шарах зростає лінійно при збільшенні розмірності моделі (розмірності векторних представлень вхідних токенів). На відміну від шарів у рекурентних та згорткових мережах, де зростання квадратичне, моделі архітектури Transformer менше чутливі до збільшення розмірності. При цьому, ситуація із впливом кількості слів у вхідній послідовності обернена, тобто обчислювальна складність із збільшенням величини контексту зростає сильніше ніж в інших моделей.

Останній недолік компенсується меншою кількістю послідовних операцій, необхідних у процесі роботи мережі. В шарах Self-Attention обчислення добре розділяються між паралельними потоками.

При наявності потужних графічних процесорів, що здатні проводити обчислення з дуже високою кількістю паралельних потоків, а також можливістю створювати кластери з таких пристроїв і горизонтально масштабувати потужності, з’явилась можливість створювати великі мовні моделі (англ. LLM – Large Language Model) із мільярдами чи навіть трильйонами параметрів. Тренуючись на терабайтах тренувальних даних, такі моделі є дуже потужним засобом для вирішення задач у сфері обробки мови.

Приклади таких моделей: GPT (OpenAI), PaLM, BERT (обидві Google), Claude (Anthropic) та LLaMA (Meta) [2]. Остання - сім’я моделей з відкритим кодом, що доступні для безкоштовного завантаження, дотренування на власних даних та використання.

Іншою перевагою є наявність одразу багатьох версій різного розміру. Таким чином можна вибрати модель з найкращим співвідношенням якості та витрат ресурсів для конкретних задач. Завдяки різноманітним оптимізаціям, таким як decoder-only структура та Grouped-query Attention [3], кожна модель показує дуже хороші результати порівняно із іншими моделями подібного розміру.

Всі базові моделі LLaMA доступні для використання на порталі HuggingFace. Також доступна велика кількість покращених для конкретних задач моделей, як від Meta так і від спільноти. Додатково можливе точне налаштування (fine-tuning) - це процес дотренування моделі на додаткових даних, для того щоб підготувати модель вузької спеціалізації.

Література

1. Vaswani, A. "Attention is all you need." Advances in Neural Information Processing Systems (2017).
2. Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." arXiv preprint arXiv:2302.13971 (2023).
3. Touvron, Hugo, et al. "Llama 2: Open foundation and fine-tuned chat models." arXiv preprint arXiv:2307.09288 (2023).