УДК 004.41

Власюк Р.Ю.

ДОСЛІДЖЕННЯ ПІДХОДІВ ДО СТВОРЕННЯ AI-ОРІЄНТОВАНИХ СИСТЕМ ГЕНЕРАЦІЇ ТЕКСТОВОГО КОНТЕНТУ

Вінницький національний технічний університет

Анотація

У статті досліджуються методи розробки систем генерації текстового контенту, орієнтованих на штучний інтелект. Від статистичних моделей до сучасних трансформерних архітектур розглядається розвиток методів генерації тексту. Проаналізовано основні принципи навчання великих мовних моделей, включаючи підготовку до навчання, удосконалення та підкріплення навчання за допомогою людського feedback. Визначено основні переваги, недоліки та сфери застосування цих систем.

Ключові слова: штучний інтелект, генерація тексту, NLP, трансформери, великі мовні моделі, глибинне навчання.

Abstract

The article examines methods for developing AI-oriented text generation systems. The evolution of text generation approaches is considered, ranging from statistical models to modern transformer-based architectures. The key principles of training large language models are analyzed, including pretraining, fine-tuning, and reinforcement learning with human feedback. The main advantages, limitations, and application areas of these systems are identified.

Keywords: artificial intelligence, text generation, NLP, transformers, large language models.

Вступ

На поточному етапі розвитку інформаційних технологій спостерігається швидке зростання обсягів текстових даних. Це означає, що потрібно створити інтелектуальні системи, які можуть автоматично створювати текстові матеріали високої якості. Електронна комерція, медіа, освіта, програмна інженерія та автоматизація бізнес-процесів є сферами, де ці системи широко застосовуються.

Методи обробки природної мови (NLP) є основою для систем генерації тексту, орієнтованих на штучний інтелект. Ці методи дозволяють моделювати мовні закономірності та створювати тексти, які є семантично узгодженими. Такі системи мають на меті створювати текст, який є змістовним, логічним і контекстно відповідним.

Основна частина

Первоначальні методи генерації текстів використовували шаблонні алгоритми та статистичні моделі, особливо n-грамні моделі, які використовували ймовірнісні залежності між словами в тексті. Такі моделі базувалися на підрахунку, наскільки часто послідовності слів з’являються в навчальних корпусах, і вони дозволяли створювати текст, вибравши найбільш ймовірне наступне слово. Але через те, що кількість параметрів постійно збільшується зі збільшенням довжини контексту, а також через те, що вони не можуть враховувати семантичні зв’язки між словами, які знаходяться на великій відстані одне від одного, їхні можливості значно обмежені. Це призвело до створення тексту, який часто був логічно непослідовним, але граматично коректним.

Впровадження нейронної мережі глибинного навчання, зокрема рекурентних нейронних мереж (RNN), пов’язане з подальшим розвитком методів генерації тексту. RNN мають здатність моделювати послідовності даних шляхом збереження прихованого стану. Це дозволило враховувати історичний контекст під час написання тексту. Покращені архітектури, як-от Long Short-Term Memory (LSTM) і Gated Recurrent Unit (GRU), частково вирішили проблеми затухання та вибуху градієнтів. Ці моделі краще працюють з довшими текстовими послідовностями завдяки механізмам «забування» та «запам’ятовування» інформації. Однак навіть вони мали високі обчислювальні витрати через послідовну природу обробки даних і мали обмеження у збереженні довгострокових залежностей, особливо при обробці великих текстових корпусів.

Упровадження трансформерної архітектури, яка значно відрізняється від попередніх методів, стало справжнім проривом у виробництві текстів. В її основі лежить механізм самоуваги, також відомий як самоувага. Цей механізм дозволяє моделі одночасно вивчати кожен елемент вхідної послідовності та визначити, як вони впливають один на одного. Це значно підвищує ефективність навчання та усуває необхідність постійної обробки даних. З цієї причини трансформери можуть краще враховувати глобальний контекст, що є важливим для створення змістовного та логічно узгодженого тексту. Масштабування моделей до великих кількостей параметрів і можливість паралельної обробки даних є ще одними перевагами.

Великі мовні моделі, також відомі як LLM, є основою для сучасних систем генерації тексту, які базуються на величезних наборах текстових даних, таких як книги, статті, веб-ресурси та інші джерела. Це дозволяє їм виконувати широкий спектр завдань без явного програмування та розвивати узагальнені мовні уявлення. Такі моделі можуть узагальнювати інформацію, вести розмову, перекладати та навіть вирішувати логічні задачі.

Створення AI-орієнтованих систем генерації текстів складається з кількох основних етапів. Модель навчається на великих неструктурованих корпусах даних шляхом передбачення наступного токена або відновлення пропущених частин тексту на етапі попереднього навчання. Це сприяє розвитку основних мовних навичок. Далі проводиться донавчання (fine-tuning) на спеціалізованих наборах даних. Це дозволяє адаптувати модель до різних завдань, таких як створення коду, обробка запитів користувачів або чат-боти. Завершальним кроком є оптимізація за допомогою методів навчання з підкріпленням із використанням людського зворотного зв'язку (Reinforcement Learning from Human Feedback, або RLHF). Мета цієї стратегії полягає в тому, щоб підвищити якість, релевантність і безпеку контенту, який було створено.

Обчислювальна інфраструктура є важливою для функціонування таких систем. Для навчання великих мовних моделей потрібні потужні графічні процесори (GPU), тензорні процесори (TPU) та розподілені обчислювальні системи. Зменшення затримок при генерації тексту, забезпечення масштабованості систем і оптимізація пам'яті також є важливими питаннями.

AI-орієнтовані системи генерації тексту мають широкий спектр застосування. Написання програмного коду, машинний переклад, аналіз і узагальнення текстів, автоматичне створення новинних статей і маркетингового контенту, створення інтелектуальних чат-ботів і віртуальних асистентів, а також адаптація контенту до потреб користувачів є частиною цих завдань. Такі системи можуть значно підвищити продуктивність роботи та скоротити витрати на обробку даних у бізнес-середовищі.

Тим не менш, є деякі проблеми з використанням цих систем. Одним із найважливіших є явище, відоме як «галюцинації», коли модель створює інформацію, яка виглядає правдоподібною, але насправді є неправдивою. Крім того, питання етичного використання є важливими; це включає запобігання упередженості у даних, забезпечення конфіденційності та нагляд за використанням створеного контенту. Крім того, великим викликом залишається залежність від значних навчальних даних, а також висока вартість розробки та обслуговування таких систем.

Висновок

Одним із найбільш перспективних напрямків розвитку сучасних інформаційних технологій є системи генерації текстового контенту, які базуються на використанні трансформерних моделей, великих мовних моделей і сучасних методів навчання. Попри значні досягнення, залишаються проблеми з якістю, мораллю та ефективністю таких систем. Удосконалення цієї галузі сприятиме створенню більш розумних і надійних систем генерації тексту.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Vaswani A. Attention is All You Nee [Електронний ресурс] – Режим доступу: https://arxiv.org/abs/1706.03762
2. Brown T. Language Models are Few-Shot Learners [Електронний ресурс] – Режим доступу: https://arxiv.org/abs/2005.14165
3. Devlin J. BERT: Pre-training of Deep Bidirectional Transformers [Електронний ресурс] – Режим доступу: https://aclanthology.org/N19-1423/
4. Goodfellow I. Deep Learning [Електронний ресурс] – Режим доступу: https: //www.deeplearningbook.org/
5. OpenAI Research Papers [Електронний ресурс] – Режим доступу: https://openai.com/research/index/publication/

Власюк Руслан Юрійович– студент групи 4ПІ-22Б, факультет інформаційних технологій і комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, e-mail: vlasiukru@gmail.com

Науковий керівник – Стахов Олексій Ярославович, доктор філософії PHD, старш.викл., Вінницький національний технічний університет, м. Вінниця, e-mail: aleksey.stahov@gmail.com

Vlasiuk Ruslan Yuriyovych – student of group 4PI-22B, Faculty of Information Technologies and Computer Engineering, Vinnytsia National Technical University, Vinnytsia, e-mail: vlasiukru@gmail.com

Supervisor – Stakhov Oleksii Yaroslavovych, PhD, Senior Lecturer at the Department of Software Engineering, Vinnytsia National Technical University, Vinnytsia, e-mail: aleksey.stahov@gmail.com