АНАЛІЗ МЕТОДІВ ГЕНЕРАЦІЇ СИНТЕТИЧНИХ ДАНИХ З ВИКОРИСТАННЯМ ГЕНЕРАТИВНОГО ШТУЧНОГО ІНТЕЛЕКТУ

Бабаніна А.О., Коваленко А.А., Ситник О.В.
Харківський національний університет радіоелектроніки, Харків, Україна

Постійне зростання досліджень, зосереджених на створенні синтетичних даних із великих мовних моделей (LLM), особливо для сценаріїв з обмеженою доступністю даних, впроваджує помітні зміни в генеративному штучному інтелекті (ШІ). Здатність використовувати ці дані на рівні з реальними даними робить цей підхід переконливим рішенням для проблем із низьким ресурсом. У цьому досліджені було розгляуто передові технології, які використовують ці гігантські LLM для генерації навчальних даних для конкретних завдань.

Метою цієї роботи є аналіз методології, методів оцінювання та практичного застосування. В роботі проаналізовано поточні обмеження та запропоновано потенційні шляхи для майбутніх досліджень.

Поява Transformer [1], а потім новаторські LLM, такі як GPT від OpenAI і BERT від Google, створили початок нової ери в розумінні та генерації мови. Нещодавно генеративні LLM розвинули цю еволюцію до нових висот, плавно поєднуючись із Generative AI і провіщаючи нову еру в сфері генерації синтетичних даних [2]. Початок Generative AI можна простежити до основних моделей, таких як Generative Adversarial Networks і Variation Autoencoders, які продемонстрували здатність створювати реалістичні зображення та сигнали. Однак справжній розвиток Generative AI почав лише з появою LLM в останні роки. Ці моделі, які навчалися на величезних наборах даних, продемонстрували безпрецедентну здатність створювати зв'язний і контекстуально релевантний текст, розсуваючи межі того, чого ШІ може досягти в завданнях, пов'язаних із мовою. Синтез Generative AI і LLM у сфері створення синтетичних даних є не просто технологічним прогресом, але й глибокою зміною парадигми в нашому підході до створення даних і навчання моделей ШІ.

В доповіді представлено деякі методи практичного навчання для навчання подальших моделей на основі синтетичних даних, припускаючи, що якість даних є невідповідною.

Список літератури

1. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023

2. Rudenko O., Bezsonov O., Vashchenko K., Rutska S. Synthetic Dataset Generation for Efficient Neural Network Training. Proceedings of the 7th International Conference on Computational Linguistics and Intelligent Systems. 2023. Т. 3387. С. 146–159.