УДК 004.8

МЕТОДИ СИНТЕЗУ НАДШВИДКОДІЮЧИХ СТРУКТУР МОВНИХ СИСТЕМ ШТУЧНОГО ІНТЕЛЕКТУ

Шульга В.В.
Науковий керівник – д.т.н., проф. Четвериков Г. Г.
Харківський національний університет радіоелектроніки, каф. ПІ
м. Харків, Україна
тел.: +38(095) 644-45-12, e-mail: vladyslav.shulha@nure.ua.

This work has goal to address the current approaches of implementing superfast language syntax parsers. It contains general overview of modern solutions based on neural networks and a more in depth description was provided for a particular model that is implemented using transformer networks – BERT. This includes providing benchmarks to understand its current capabilities and field where it can be possibly used.

Синтез мовних структур займає вагоме місце серед сучасних програмних систем, в тому числі тих, які базуються на штучному інтелекті. Впродовж останніх років, проблематика реалізації систем із реалізацією цих систем для набула великої популярності, не в останню чергу, через вибухове зростання даних і необхідності швидкої та ефективної обробки природної мови. Високошвидкісні структури є важливим аспектом цих систем, за допомогою них, штучний інтелект має можливість до швидкої взаємодії зі своїми користувачам, впливаючи на їх життя та працю. Практичне відображення результату роботи цих систем, прямо зараз ми можемо побачити на прикладах чат-асистентів, які дозволяють машинам розуміти та інтерпретувати людську мову, уможливлюючи взаємодію з людьми більш природним та інтуїтивно зрозумілим способом.

Однак синтез мовних структур залишається складною проблемою. Однією з ключових проблем є потреба у великих обсягах високоякісних даних для навчання цих систем, які може бути важко отримати. Крім того, існує потреба в алгоритмах, які можуть ефективно обробляти ці дані та вивчати складні шаблони природною мовою. Досить високої гостроти також має проблема з розміщенням моделей на обчислювальних машинах, на які сформувався дефіцит через ріст галузі, і відповідно з`явилися обмеження до ресурсів.

Існують різні методи реалізації синтезу високошвидкісних структур у мовних системах штучного інтелекту. Одним із поширених підходів є використання апаратного прискорення для прискорення обробки даних природної мови. Це передбачає використання спеціалізованого апаратного забезпечення, такого як графічні процесори (GPU) або програмовані вентильні матриці (FPGA), для прискорення виконання моделей нейронних мереж. Інший підхід полягає в оптимізації програмної реалізації алгоритмів обробки природної мови. Це включає такі методи, як

скорочення, квантування та розподіл ваги, щоб зменшити обчислювальну складність моделей нейронних мереж, зберігаючи при цьому точність.

До прикладів алгоритмічної реалізації обробки даних природної мови, можна віднести наступні з відносно розповсюджених підходів, на яких базується розробка:

Існує три метода контролю положення сонячних панелей:

– згорткові нейронні мережі (CNN);

– рекурентні нейронні мережі (RNN);

– моделі на основі трансформаторних мереж.

До останньої категорії також відноситься BERT (Bidirectional Encoder Representations from Transformers або двоспрямовані кодувальні представлення з трансформерів) – це попередньо навчена мовна модель, розроблена Google, яка досягла високого рівня продуктивності в широкому діапазоні завдань обробки природної мови. Успіх BERT частково пояснюється його двонаправленою архітектурою, яка дозволяє отримувати глибоку контекстну інформацію про слова в реченні.

Переходячи до детального розгляду результатів, BERT досяг гарних результатів у кількох контрольних наборах даних, включаючи Stanford Question Answering Dataset (SQuAD) і загальне оцінювання розуміння мови (GLUE). На SQuAD BERT досяг результату F1 у 93,2%, перевершивши попередній сучасний результат у 90,9%. У тесті GLUE BERT перевершив усі інші моделі, досягнувши середнього результату 80,5%, що на 7,7% вище, ніж попередній найсучасніший результат.

Загалом, ефективність BERT у широкому діапазоні завдань обробки природної мови затвердила її як одну з найбільш впливових мовних моделей у цій галузі. Його успіх мав вплив на подальші дослідження нових методів і архітектур обробки природної мови з кінцевою метою розробки ще точніших і ефективніших мовних моделей.

Список використаних джерел:

1. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1810.04805

2. Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, Samuel R. Bowman (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. https://arxiv.org/abs/1804.07461

3. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://arxiv.org/abs/1907.11692

4. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A Lite BERT for Self-supervised Learning of Language https://arxiv.org/abs/1909.11942