УДК 004.85

РОЗРОБКА КОМПАКТНОЇ МОДЕЛІ МАШИННОГО ПЕРЕКЛАДУ В УМОВАХ ОБМЕЖЕНИХ ОБЧИСЛЮВАЛЬНИХ РЕСУРСІВ

Кутько В.О.
e-mail: vladyslav.kutko@nure.ua
Харківський національний університет радіоелектроніки, каф. СТ
м. Харків, Україна

The development of compact machine translation models for resource constrained environments is a pressing challenge in modern computational linguistics. This study focuses on creating an efficient translation system between English and Ukrainian, optimized for real-time performance on low-power devices. The proposed model, based on a transformer architecture with up to 100 million parameters. Utilizing PyTorch, quantization, and pruning techniques, the system is designed for autonomous applications, such as mobile apps and voice assistants. Comparative analysis with models like OPUS-MT highlights its competitive edge in compactness and adaptability.

Сучасні системи машинного перекладу, побудовані на базі архітектур трансформерів [1], демонструють високу якість перекладу, однак їхнє використання в умовах обмежених обчислювальних ресурсів залишається проблематичним через велику кількість параметрів і високі вимоги до апаратного забезпечення. Розробка компактних моделей, здатних працювати в реальному часі на пристроях із низькою потужністю, таких як смартфони чи вбудовані системи, є актуальним завданням. Метою роботи є створення ефективної системи машинного перекладу між англійською та українською мовами, яка забезпечує високу якість перекладу при мінімальних затримках, низькому споживанні оперативної пам’яті (RAM) та енерговитрат.

Задача машинного перекладу в умовах обмежених ресурсів вимагає балансу між якістю перекладу та продуктивністю системи. У роботі розглянуто двосторонній переклад із акцентом на адаптацію до специфіки української мови [2]. Система має відповідати таким вимогам: інференс у реальному часі (затримка до 1 с), споживання RAM до 1–2 ГБ. Для досягнення цих цілей використано методи оптимізації, такі як квантова апроксимація та прунинг, які зменшують розмір моделі до 100 млн параметрів, зберігаючи якість перекладу.

Аналіз існуючих рішень показав, що моделі на базі трансформерів, такі як OPUS-MT (BLEU 32.20) та Dragoman PT (BLEU 32.34), є конкурентоспроможними, але мають обмеження: перша застаріла (2020 рік), а друга не підтримує зворотний переклад. Розроблювана модель перевершує їх за компактністю та адаптивністю, що дозволяє застосовувати її в автономних додатках (мобільні застосунки, голосові асистенти) та спеціалізованих доменах (юридичні, медичні тексти).

Система розроблена на базі фреймворку PyTorch, який забезпечує гнучкість у створенні нейронних мереж та оптимізацію для GPU-обчислень. Середовищем розробки обрано Cursor завдяки підтримці автодоповнення та інтеграції з Git. Обчислювальна платформа – NVIDIA RTX 4070 Ti Super із 16 ГБ відеопам'яті – дозволяє ефективно тренувати та тестувати модель. Для оптимізації використано бібліотеки torch.quantization та TensorRT, які зменшують розмір моделі та прискорюють інференс. Якість перекладу оцінюється метрикою BLEU, цільовий показник – вище 32. Модель базується на архітектурі трансформерів із модульною структурою [3]:

– попередня обробка (токенізація, нормалізація);

– ядро перекладу (інференс);

– постобробка (корекція граматики);

– інтерфейс користувача.

Розроблена модель демонструє можливість перекладу в реальному часі з затримкою до 1 с та споживанням RAM до 1–2 ГБ, що робить її придатною для пристроїв із низькою потужністю. Адаптація до української мови підвищує точність перекладу складних конструкцій, а модульність дозволяє налаштування під різні домени. Порівняно з універсальними моделями (наприклад, GPT-4), система має перевагу в ефективності та спеціалізації.

Розроблена компактна модель машинного перекладу поєднує високу якість перекладу з оптимізованим використанням ресурсів, що робить її перспективною для впровадження в автономних системах реального часу. Подальші дослідження можуть бути спрямовані на підвищення якості перекладу для спеціалізованих доменів та інтеграцію з IoT-пристроями.

Список використаних джерел:

1. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I. Attention Is All You Need // Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 5998–6008.

2. Kumar S., Anastasopoulos A., Wintner S., Tsvetkov Y. Machine Translation into Low-resource Language Varieties // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). 2021. P. 110–121.

3. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. Vol. 1. P. 4171–4186.