КВАНТУВАННЯ НЕЙРОМЕРЕЖЕВИХ СИСТЕМ РОЗПІЗНАВАННЯ ГОЛОСУ

Сальніков Д.В.
Національний технічний університет «Харківський політехнічний інститут», м. Харків

На поточний момент широкого розповсюдження набули автоматизовані системи що використовують методи штучного інтелекту та нейронні мережі. Такі системи мають високі характеристики якості роботи та дозволяють оброблювати широкий спектр різноманітних даних. Зокрема розпізнавати голосові команди. Одними з найбільш поширених систем розпізнавання голосу людини можна вважати Wav2vec [1] та Wisper [2].

Якість роботи подібних алгоритмів забезпечується великою кількістю слоїв нейронної мережі та, як наслідок, великою кількістю параметрів навчання. Такі параметри не змінюються під час експлуатації системи але накладають дуже високі обмеження на обчислювальну систему на якій відбуваються розрахунки. Ці обмеження не дозволяють використовувати такі моделі у сучасних вбудованих та мікропроцесорних системах, що знижує можливість їх використання в системах з обмеженою можливістю підключення до мережі.

Для обмеження об'єму необхідної пам'яті вагові коефіцієнти нейронних мереж квантують. В роботі [3] наведено механізми квантування великих мовних моделей до низької розрядності. Залежно від налаштувань такий підхід дозволяє знизити використання памя'яті в 2 – 5 разів для моделей LLaMA LLM, та прискорити виконання моделі в 1.2 – 3 рази.

Виходячи із сказаного стоїть задача квантування моделей Wav2Vec та Wisper до низьких розрядностей параметрів для істотного зниження необхідної пам'яті та спрощення розрахункової складності моделей. В ході роботи виконано реалізацію моделі Wav2Vec мовою Python із використанням бібліотеки numpy для проведення моделювання впливу різних стратегій квантування на якість роботи моделі. Для контролю яксоті та розрахункової складності розпізнавання використовується набір даних LibriSpeech.

Література:

1. Schneider, S., Baevski, A., Collobert, R., Auli, M. (2019) wav2vec: Unsupervised Pre-Training for Speech Recognition. Proc. Interspeech 2019, 3465-3469, doi: 10.21437/Interspeech.2019-1873

2. A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large- Scale Weak Supervision," 2022. [Online]. Available: https: //arxiv.org/abs/2212.04356

3. Ma, S., Wang, H., Ma, L., Wang, L., Wang, W., Huang, S., Dong, L., Wang, R., Xue, J., & Wei, F. (2024). The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits. ArXiv, abs/2402.17764.