ОПТИМІЗАЦІЯ НЕЙРОННИХ МЕРЕЖ В УМОВАХ ОБМЕЖЕНИХ ОБЧИСЛЮВАЛЬНИХ РЕСУРСІВ

магістр Д.А. Стецюк, канд. техн. наук, доц. В.М. Савченко,
ст. викладач О.В. Мнушка, НТУ "ХПІ", м. Харків

Пристрої з обмеженими обчислювальними можливостями використовують для побудови сучасних систем на основі Інтернету речей та промислового Інтернету речей [1 – 3]. Впровадження у такі системи рішень на основі штучного інтелекту (ШІ) потребує масового застосування глибоких нейронних мереж (НМ), при цьому розробники стикаються із проблемою адаптації надпотужних НМ для подібних систем [4 – 5], в першу чергу із-за жорстких обмежень на споживання енергії, обчислювальну потужність та доступний обсяг пам’яті.

Задачею дослідження є розробка автоматизованого програмного засобу для ефективної оптимізації попередньо навчених НМ для їхнього подальшого використання в умовах обмежених ресурсів.

Практична значимість роботи полягає в створенні уніфікованого інструменту, який поєднує та систематизує різні підходи до оптимізації, надає порівняльну оцінку їх ефективності для конкретної моделі та цілей користувача (наприклад, максимальне стиснення або максимальна швидкість), а також пропонує рекомендації щодо вибору методу.

Отримані результати підтверджують гіпотезу про те, що комбіноване застосування сучасних методів оптимізації дозволяє досягти високого коефіцієнта стиснення та прискорення при визначених втратах якості. Розроблений програмний засіб довів свою ефективність і може бути корисним під час розгортанні нейронних мереж на пристроях з обмеженими обчислювальними ресурсами.

Список літератури: 1. *Mnushka O.V.* Edge computing for solutions based on the Internet of Things / O.V. Mnushka, V.M. Savchenko // Information technologies: science, engineering, technology, education, health. Scientific publication. Abstracts. XXVIII International scientific-practical conference MicroCAD-2020. P. IV. Kharkiv, 2020. – P. 179. **2.** *Mnushka O.* Continuous integration for a development process of the information technology of remote monitoring and control / O. Mnushka, S. Leonov, V. Savchenko // Herald of NTU "KhPI". Series: Informatics and modeling. – Kharkov: NTU "KhPI". - 2022. - № 1 – 2 (7 – 8). - P.5-17. DOI: 10.20998/2411-0558.2022.02.01. **3.** *Мнушка О.В.* SCADA на основі промислового Інтернету речей: архітектура системи / О.В. Мнушка // Технічний сервіс агропромислового, лісового та транспортного комплексів. – Харків, 2018. – №12. – С.117-124. – ISSN 2311-441X. ***4.*** Giovannesi L. OptDNN: Automatic deep neural networks optimizer for edge computing / L. Giovannesi, G. Proietti Mattia, R. Beraldi // Software Impacts. – 2024. – Vol. 20. – P. 100641. – ISSN 2665-9638. – DOI: 10.1016/j.simpa.2024.100641. **5.** Zakovorotnyi O. Optimization of Neural Network Calculations Using Integer Arithmetic / O. Zakovorotnyi, A. Khulap // 2024 IEEE 5th KhPI Week on Advanced Technology (KhPIWeek). – Kharkiv, Ukraine, 2024. – P. 1–4. – DOI: 10.1109/KhPIWeek61434.2024.10877949..