Облікова картка дисертації (ОКД)

Шифр спецради: ДФ 26.002.05

Відкрита

Вид дисертації: 08

Державний обліковий номер: 0823U100075

Дата реєстрації: 13-02-2023

1. Відомості про здобувача

ПІБ (укр.): Таран Владислав Ігорович

ПІБ (англ.): Taran Vladyslav I.

Шифр спеціальності, за якою відбувся захист: 123

Дата захисту: 08-02-2023

На здобуття наукового ступеня: Доктор філософії (д.філ)

Спеціальність за освітою: Комп'ютерна інженерія

2. Відомості про установу, організацію, у вченій раді якої відбувся захист

Назва організації: Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"

Підпорядкованість: Міністерство освіти і науки України

Код ЄДРПОУ: 02070921

Адреса: проспект Перемоги, буд. 37, м. Київ, 03056, Україна

Телефон: 380442367989

Телефон: 380442044862

Телефон: +38 (044) 204-82-82

E-mail: mail@kpi.ua

WWW: https://kpi.ua/

Інше: kpi.ua

3. Відомості про організацію, де виконувалася (готувалася) дисертація

Назва організації: Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"

Підпорядкованість: Міністерство освіти і науки України

Код ЄДРПОУ: 02070921

Адреса: проспект Перемоги, буд. 37, м. Київ, 03056, Україна

Телефон: 380442367989

Телефон: 380442044862

Телефон: +38 (044) 204-82-82

E-mail: mail@kpi.ua

WWW: https://kpi.ua/

Інше: kpi.ua

4. Відомості про організацію, де працює здобувач

Назва організації: Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"

Підпорядкованість: Міністерство освіти і науки України

Код ЄДРПОУ: 02070921

Адреса: проспект Перемоги, буд. 37, м. Київ, 03056, Україна

Телефон: 380442367989

Телефон: 380442044862

Телефон: +38 (044) 204-82-82

E-mail: mail@kpi.ua

WWW: https://kpi.ua/

Інше: kpi.ua

5. Наукові керівники та консультанти

Наукові керівники

Гордієнко Юрій Григорович (д.ф.-м.н., с.н.с., 01.04.13)

6. Офіційні опоненти та рецензенти

Офіційні опоненти

Терещенко Василь Миколайович (д.ф.-м.н., професор, 01.05.01)

Тульчинський Вадим Григорович (д.ф.-м.н., с.н.с., 01.05.03)

Рецензенти

Терейковський Ігор Анатолійович (д. т. н., професор, 05.13.21)

Стеценко Інна Вячеславівна (д.т.н., професор, 05.13.06)

7. Підсумки дослідження та кількісні показники

Підсумки дослідження: 40 - Нове вирішення актуального наукового завдання
Кількість сторінок: 152
Кількість додатків: 2
Ілюстрації: 58
Таблиці: 6
Схеми:
Використані першоджерела: 118

Кількість публікацій: 12
Кількість патентів:
Впровадження результатів роботи:
Мова документа: Українська
Зв'язок з науковими темами:

8. Індекс УДК тематичних рубрик НТІ

Індекс УДК: 004.8.032.26, 004.032.26 (043.3)

Тематичні рубрики: 28.23.37

9. Тема та реферат дисертації

Тема (укр.)

Метод адаптації глибоких нейронних мереж до апаратного забезпечення зі спеціалізованою архітектурою

Тема (англ.)

Method of deep neural networks adaptation for hardware with specialized architecture

Реферат (укр.)

Дисертаційна робота присвячена розробці комплексного методу адаптації глибоких нейронних мереж, що дозволяє підвищити продуктивність та ефективність обробки даних глибокими нейронними мережами на апаратному забезпеченні зі спеціалізованою архітектурою. Вперше було розроблено комплексний метод адаптації глибоких нейронних мереж для спеціалізованих обчислювальних архітектур. Розроблено метод адаптивного ітеративного прунінгу для зменшення розміру моделей нейронних мереж за рахунок поступового зменшення розміру мережі шляхом видалення зайвих каналів у згорткових шарах та додатковому навчанні отриманої зменшеної моделі для відновлення точності розпізнавання. Відповідно до розробленого методу, гіперпараметри мережі адаптивно змінюються, щоб компенсувати втрати точності після кожної ітерації прунінгу та зменшити час ітерації обробки даних. Розроблено метод підвищення ефективності процесу обробки даних нейронними мережами на спеціалізованих обчислювальних архітектурах, що враховує технічні особливості обробки даних за допомогою глибоких нейронних мереж на спеціалізованих прискорювачах, наприклад, ітерація виконання обчислень. Також цей метод дозволяє визначити параметри такі, як розмір порції даних, щоб збільшити продуктивність обробки даних за рахунок зменшення впливу накладних витрат ініціалізації і передачі даних. Розроблено метод підвищення ефективності інфраструктури для обробки даних за допомогою глибоких нейронних мереж за рахунок зміни програмної та апаратної складової такої, як операційна система та інтерфейси підключення. Це дозволяє збільшити продуктивність та ефективність обробки даних за допомогою нейронних мереж на цільовій системі. Розроблено програмний компонент діагностики легеневих аномалій за даними рентген знімків для дослідження ефективності роботи спеціалізованого прискорювача Coral Edge TPU USB в задачах медичного застосунку. В якості архітектури глибокої нейронної мережі для даної задачі було обрано ResNet50, яку було треновано на наборі даних ChestXray та адаптовано під спеціалізований прискорювач відповідно до розробленого комплексного методу адаптації. Проведено аналіз результатів застосування методу адаптації глибоких нейронних мереж, що включає в себе адаптивний ітеративний прунінг, підвищення ефективності процесу обробки даних нейронною мережею та підвищення ефективності програмно-апаратної складової цільової хост-системи. За результатами застосування розробленого методу адаптивного ітеративного прунінгу було досягнуто прискорення 32,2 із точністю розпізнавання 96,2% (10 ітерацій прунінгу). За результатами аналізу технічних особливостей роботи спеціалізованих обчислювальних архітектур було виявлено, що значні показники прискорення, при використанні TPU у порівнянні з GPU, досягаються на пізніх ітераціях (>3) виконання обробки даних моделями глибоких нейронних мереж, коли витрати на ініціалізацію не впливають на продуктивність. Даний фактор треба враховувати при підвищенні ефективності процесу обробки даних нейронними мережами на прискорювачах зі спеціалізованою архітектурою. В результаті аналізу факторів, що впливають на продуктивність цільової інфраструктури обробки даних за допомогою глибоких нейронних мереж, було досягнуто значних різниць в продуктивності при застосуванні різних комбінацій забезпечення цільової інфраструктури. При цьому, досягнуте прискорення склало 8,7. Розроблені методи є складовою комплексного методу адаптації глибоких нейронних мереж і дозволяють підготувати обрану модель нейронної мережі для її застосування на зазначених вище прискорювачах нейронних мереж зі спеціалізованою архітектурою.

Реферат (англ.)

Thesis is devoted to the development of the complex adaptation method of deep neural networks, which allows to increase productivity and efficiency of deep neural networks applications on the hardware with specialized architecture. The complex deep neural networks adaptation method for specialized hardware was developed. The method of adaptive iterative pruning for decreasing neural network model size was developed, which is based on subsequent decrease of the model size by removing redundant channels in convolution layers and additional model training for accuracy recovery. According to the proposed method, model hyper parameters are changed after every iteration to compensate accuracy loss and to achieve time decreasing of data processing iteration. The method of neural network data processing efficiency improvement for specialized accelerators was developed. It is based on the technical aspects of deep neural network data processing on hardware with specialized

architectures, for example data processing iteration and allows to determine processing parameters for decreasing influence of overheads. The method of neural network processing infrastructure efficiency improvement was developed. It allows to optimize hardware and software configuration of the target system for increasing deep neural network data processing productivity. The testing software for medical diagnostics in the context of edge computing was developed. It utilizes the developed deep neural networks adaptation method and specialized accelerator Coral Edge TPU. The result analysis of the deep neural network adaptation method application was performed. It includes adaptive iterative pruning method, data processing efficiency improvement method and computational infrastructure efficiency improvement method. The speedup up to 32,2 and 96,2% accuracy were achieved after performing 10 iteration of the developed adaptive iterative pruning method. Based on the technical processing properties analysis for specialized processing architectures, some factors were identified, which have influence on the data processing. The considerable speedup values, while utilizing TPU compared to GPU, were achieved on the later data processing iterations (>3) with deep neural networks models, when initialization overheads had small influence on the accelerator performance. Such factor should be taken into account, while improving deep neural networks data processing efficiency on the accelerators with specialized architecture. Based on the deep neural network processing infrastructure analysis of factors, which had influence on the processing productivity, the following was identified. Considerable productivity difference was achieved, while utilizing different software and hardware combinations of the processing infrastructure. The achieved speedup value was up to 8,7. Developed methods are parts of the complex deep neural networks adaptation method. It allows to prepare the selected neural network model for application on the accelerator with specialized architecture.

Голова спеціалізованої вченої ради: Новотарський Михайло Анатолійович (д.т.н., с.н.с., 01.05.02)

Підпис

М.П.

Відповідальний за подання документів: Таран В.І. (Тел.: 0953177280)

Підпис

Керівник відділу реєстрації наукової діяльності УкрІНТЕІ Юрченко Т.А.