Облікова картка дисертації

I. Загальні відомості

Державний обліковий номер: 0824U001849

Особливі позначки: відкрита

Дата реєстрації: 08-05-2024

Статус: Наказ про видачу диплома

Реквізити наказу МОН / наказу закладу: № НСВС_62_24 від 23.07.2024

II. Відомості про здобувача

Власне Прізвище Ім'я По-батькові:

1. Мельниченко Артем Васильович
2. Artem V. Melnychenko

Кваліфікація:

Ідентифікатор ORCID ID: 0009-0000-3588-4772

Вид дисертації: доктор філософії

Аспірантура/Докторантура: так

Шифр наукової спеціальності: 121

Назва наукової спеціальності: Інженерія програмного забезпечення

Галузь / галузі знань:

Освітньо-наукова програма зі спеціальності: Інженерія програмного забезпечення

Дата захисту: 03-07-2024

Спеціальність за освітою: Інженерія програмного забезпечення

Місце роботи здобувача:

Код за ЄДРПОУ:

Місцезнаходження:

Форма власності:

Сфера управління:

Ідентифікатор ROR: Не застосовується

III. Відомості про організацію, де відбувся захист

Шифр спеціалізованої вченої ради (разової спеціалізованої вченої ради): ДФ 26.002.171; ID 5617

Повне найменування юридичної особи: Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"

Код за ЄДРПОУ: 02070921

Місцезнаходження: проспект Берестейський, буд. 37, Київ, 03056, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

IV. Відомості про підприємство, установу, організацію, в якій було виконано дисертацію

Повне найменування юридичної особи: Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"

Код за ЄДРПОУ: 02070921

Місцезнаходження: проспект Берестейський, буд. 37, Київ, 03056, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

V. Відомості про дисертацію

Мова дисертації: Українська

Коди тематичних рубрик: 28.23.15, 28.23.37, 20.54.07

Тема дисертації:

1. Методи та програмні засоби підвищення швидкодії моделей розпізнавання образів на основі машинного навчання

2. Methods and software tools for improving the performance of pattern recognition models based on machine learning

Реферат:

1. Дисертаційна робота присвячена аналізу методів оптимізації нейронних мереж і розробці програмних засобів для збільшення швидкодії нейронних мереж під час навчання і виконання. У сучасному високотехнологічному світі, нейронні мережі вийшли на передній план як ключова технологія. Ця варіація математичних моделей продемонструвала високу ефективність у багатьох задачах, що варіюються від комп'ютерного зору до розуміння природніх мов, тим самим ставши невід'ємною частиною щоденного життя. Втім, розгортання нейронних мереж у реальних сценаріях часто ускладняється їхньою обчислювальною складністю та ресурсоємністю. Великий об’єм енергоспоживання, що потребується для

навчання і використання великих моделей нейронних також має негативний вплив на навколишнє середовище. Обчислювальна складність часто проявляється у вигляді великої кількості параметрів та глибоких архітектур, які вимагають значного об’єму обчислювальної потужності як для навчання, так і для подальшого використання на кінцевих пристроях. Ця складність є особливо проблематичною в застосуваннях нейронних мереж на пристроях Інтернету речей (IoT), де обчислювальні ресурси часто обмежені. Ресурсоємні характеристики включають в себе обчислювальну потужність і використання пам'яті. Це питання є особливо актуальним у мобільних та вбудованих пристроях, де пам'ять є обмеженим ресурсом. Більше того, затримка, спричинена нестачею ресурсів, часто є неприйнятною в ряді задач, що включає в себе системи автономного керування, де навіть невелика затримка в прийнятті рішень може мати серйозні наслідки. Оптимізація нейронних мереж є актуальною задачею в технологічній галузі, що підкреслюється емпіричними даними. Об’єм обчислювальних ресурсів, необхідний для навчання найсучасніших нейронних мереж, подвоювався приблизно кожні 3 місяці з 2012 року. Це експоненційне зростання обчислювальних вимог не є сталим на довгострокову перспективу, особливо з урахуванням енергоспоживання та екологічного впливу, пов'язаного з дата-центрами. Метою дисертації є збільшення ефективності моделей нейронних мереж, а саме зменшення втрати точності при збільшенні швидкодії, після застосування методів оптимізації моделей глибинного навчання, створених для вирішення задач комп’ютерного зору. Наукова новизна одержаних результатів полягає в наступному. Удосконалено модель нейронної мережі для виявлення облич RetinaFace, яка на відміну від існуючих використовує метод прунінгу SNIP для оптимізації, що дозволяє використовувати розріджені матриці для зберігання і виконання мережі з метою подальшого удосконалення та збільшення швидкодії. Удосконалено метод прунінгу SNIP для моделі виявлення облич RetinaFace, який на відміну від існуючих передбачає можливість виключення контекстних модулів з процесу прунінгу. Вдосконалений метод дозволяє досягти більшої точності при незмінній кількості виключених параметрів. Вперше розроблено метод прунінгу перед навчанням для моделей архітектури трансформер, який на відміну від існуючих враховує важливість механізму «уваги». Використання розробленого методу дозволяє значно збільшити точність класифікації кінцевої моделі в порівнянні з методом SNIP. Вперше розроблено архітектуру програмного забезпечення для моделювання та дослідження методів прунінгу перед навчанням нейронних мереж, яка на відміну від існуючих дозволяє приводити матриці вагових коефіцієнтів мережі до розрідженого формату, використовуючи запропонований механізм оцінки важливості вагів. Оптимізована мережа RetinaFace містить на 68% параметрів менше ніж початкова мережа при втраті точності на лише 1.4%. Вдосконалений метод дозволив зменшити втрати точності з 1.4% до 0.7% порівняно з методом SNIP при порівнянні з необрізаною моделлю, при скороченні параметрів на 68%. Реалізація методу прунінгу для архітектури трансформер дозволила натренувати мережу з покращенням точності до 37% порівняно з методом SNIP при порівнянні з необрізаною моделлю, при скороченні кількості параметрів на 90%. Встановлено, що результати визначення критеріїв важливості вагів, отриманих розробленим алгоритмом, можуть бути використані для підвищення швидкодії нейронних мереж від 20% до 65% шляхом використання розріджених матриць формату 2:4, в залежності від графічного процесора. Встановлено, що додаткові виходи для сіамських нейронних мереж, призначених для встановлення схожості двох зображень, не дають приросту в швидкості сходження і точності моделі.

2. This dissertation is devoted to the analysis of neural network optimization methods and the development of software tools to increase the performance of neural networks during training and execution. In today's high-tech world, neural networks have come to the forefront as a key technology. This variation of mathematical models has demonstrated high performance in many tasks ranging from computer vision to natural language understanding, thereby becoming an integral part of everyday life. However, the deployment of neural networks in real-world scenarios is often hampered by their computational complexity and resource intensity. The large amount of power consumption required to train and use large neural models also has a negative impact on the environment. Computational complexity often manifests itself in the form of a large number of parameters and deep architectures that require a significant amount of computing power both for training and for further use on end devices. This complexity is particularly problematic in applications of neural networks on Internet of Things (IoT)

devices, where computing resources are often limited. Resource-intensive characteristics include computing power and memory usage. This issue is particularly relevant in mobile and embedded devices where memory is a limited resource. Moreover, the latency caused by a lack of resources is often unacceptable in a number of tasks, including autonomous control systems, where even a small delay in decision-making can have serious consequences. Optimization of neural networks is an urgent task in the technology industry, which is emphasized by empirical data. The amount of computing resources required to train state-of-the-art neural networks has doubled approximately every 3 months since 2012. This exponential growth in computational requirements is not sustainable in the long run, especially considering the energy consumption and environmental impact associated with data centers. The purpose of this thesis is to increase the efficiency of neural network models, namely, to reduce the loss of accuracy while increasing performance, after applying methods to optimize deep learning models created to solve computer vision problems. The scientific novelty of the results is as follows. An improved model of the RetinaFace neural network for face detection is proposed, which, unlike the existing ones, uses the SNIP pruning method for optimization, which allows the use of sparse matrices for storing and executing the network for further improvement and performance. An improved SNIP pinning method for the RetinaFace face detection model is proposed, which, unlike the existing ones, provides for the possibility of excluding contextual modules from the pinning process. The improved method allows achieving higher accuracy with the same number of excluded parameters. For the first time, a pre-training tuning method for transformer architecture models has been developed, which, unlike the existing ones, takes into account the importance of the "attention" mechanism. The use of the developed method allows to significantly increase the accuracy of classification of the final model compared to the SNIP method. For the first time, a software architecture for modelling and studying pre-training methods for neural networks has been developed, which, unlike existing ones, allows to reduce the matrix of network weights to a sparse format using the proposed mechanism for assessing the importance of weights. The optimized RetinaFace network contains 68% fewer parameters than the original network, with a loss of accuracy of only 1.4%. The improved method reduced the accuracy loss from 1.4% to 0.7% compared to the SNIP method when compared to the uncropped model, with a 68% reduction in parameters. Implementation of the pruning method for the transformer architecture allowed to train the network with an accuracy improvement of up to 37% compared to the SNIP method when compared to the uncut model, while reducing the number of parameters by 90%. The results of determining the criteria for the importance of weights obtained by the developed algorithm can be used to increase the performance of neural networks from 20% to 65% by using sparse matrices of 2:4 format, depending on the GPU. The study established that additional outputs for Siamese neural networks designed to establish the similarity of two images do not increase the speed of convergence and model accuracy.

Державний реєстраційний номер ДіР:

Пріоритетний напрям розвитку науки і техніки: Інформаційні та комунікаційні технології

Стратегічний пріоритетний напрям інноваційної діяльності: Розвиток сучасних інформаційних, комунікаційних технологій, робототехніки

Підсумки дослідження: Нове вирішення актуального наукового завдання

Публікації:

- Melnychenko, A., Zdor K. Incorporating attention score to improve foresight pruning on transformer models. Computer Science and Applied Mathematics, 2023, №2, pp.22-28
- Melnychenko, A., Shaldenko, O. Evaluation of a snip pruning method for a state-of-the-art face detection model. Computational Problems of Electrical Engineering, 2023, Vol. 12, №1, pp. 18-22
- Melnychenko, A., Zdor, K. Efficiency of supplementary outputs in siamese neural networks. Advanced Information Systems, 2023,Volume 7, №3, pp. 49–53
- Мельниченко, А., Шалденко, О. Особливості використання прунінгу перед тренуванням нейронної мережі для детекції обличчя, XX Міжнародна науково-практична конференція молодих вчених і

студентів, 25п28 квітня 2023 року, Київ, Україна

- Melnychenko A. Evaluating SNIP pruning method on the state-of-the-art face detection model. Modern scientific research: achievements, innovations and development prospects, XVI Міжнародна науково-практична конференція, 11-13 вересня 2022 року, Берлін, Німеччина. С. 68-72.5. Melnychenko A. Evaluating SNIP pruning method on the state-of-the-art face detection model. Modern scientific research: achievements, innovations and development prospects, XVI Міжнародна науково-практична конференція, 11-13 вересня 2022 року, Берлін, Німеччина. С. 68-72.
- Melnychenko, A., Zdor, K. Applying classification and regression supplemetary output in siamese neural network using fashion MNIST and plantvillage datasets, VII Міжнародна науково-практична конференція "Modern problems of science, education and society", 11-13 вересня 2023 Київ, Україна, С. 126-129.
- Melnychenko, A., & Zdor, K. Appling classification and regression supplemetary outputs in siamese neural network using plantvillage dataset, I Міжнародна науково-практична конференція "Current challenges of science and education", 18-20 вересня 2023, Берлін, Німеччина. С. 79-82.
- Melnychenko A., Zdor K. Appling classification and regression supplemetary output in siamese neural network using fashion MNIST and plantvillage datasets, X Міжнародна науково-практична конференція "Innovations and prospects in modern science", 25-27 вересня 2023, Стокгольм, Швеція. С. 87-92.
- Мельниченко А., Здор К. Збільшення ефективності оптимізації моделей архітектури ViT перед навчанням шляхом включення активацій механізму самоуваги, I міжнародна науково–практична конференція "Сучасні аспекти інженерії програмного забезпечення", 14 грудня 2023, Київ, Україна.
- Мельниченко А.В., Здор К.А. Врахування механізмів самоуваги при прунінгу моделей нейронних мереж Vision Transformer. Збірник матеріалів III Міжнародної науково-технічної конференції "Системи і технології зв'язку, інформатизації та кібербезпеки: актуальні питання і тенденції розвитку", 30 листопада 2023 року, Київ, Україна. С. 214 – 215.

Наукова (науково-технічна) продукція: програмні продукти, програмно-технологічна документація

Соціально-економічна спрямованість: економія енергоресурсів; підвищення автоматизації виробничих процесів

Охоронні документи на ОПІВ:

Впровадження результатів дисертації: Впроваджено

Зв'язок з науковими темами: 0121U109207

VI. Відомості про наукового керівника/керівників (консультанта)

Власне Прізвище Ім'я По-батькові:

1. Недашківський Олексій Леонідович

2. Oleksii Nedashkivskyi

Кваліфікація: д. т. н., професор, 05.12.02

Ідентифікатор ORCID ID: 0000-0002-1788-4434

Додаткова інформація:

Повне найменування юридичної особи: Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"

Код за ЄДРПОУ: 02070921

Місцезнаходження: проспект Берестейський, буд. 37, Київ, 03056, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

Власне Прізвище Ім'я По-батькові:

1. Шалденко Олексій Вікторович
2. Oleksii V. Shaldenko

Кваліфікація: к. т. н., доцент, 01.02.05

Ідентифікатор ORCID ID: 0000-0001-6730-965X

Додаткова інформація:

Повне найменування юридичної особи: Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"

Код за ЄДРПОУ: 02070921

Місцезнаходження: проспект Берестейський, буд. 37, Київ, 03056, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

VII. Відомості про офіційних опонентів та рецензентів

Офіційні опоненти

Власне Прізвище Ім'я По-батькові:

1. Семенов Сергій Геннадійович
2. Serhii H. Semenov

Кваліфікація: д. т. н., професор, 05.13.05

Ідентифікатор ORCID ID: 0000-0003-4472-9234

Додаткова інформація:

Повне найменування юридичної особи: Приватна установа "Університет науки, підприємництва та технологій"

Код за ЄДРПОУ: 44435841

Місцезнаходження: вул. Шпака Миколи, буд 3, Київ, 03113, Україна

Форма власності: Приватна/недержавна

Сфера управління:

Ідентифікатор ROR:

Власне Прізвище Ім'я По-батькові:

1. Сторчак Каміла Павлівна

2. Kamila P. Storchak

Кваліфікація: д. т. н., професор, 05.13.06

Ідентифікатор ORCID ID: 0000-0001-9295-4685

Додаткова інформація:

Повне найменування юридичної особи: Державний університет інформаційно-комунікаційних технологій

Код за ЄДРПОУ: 38855349

Місцезнаходження: вул. Солом'янська, буд. 7, Київ, 03110, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

Рецензенти

Власне Прізвище Ім'я По-батькові:

1. Коваль Олександр Васильович

2. Oleksandr Koval

Кваліфікація: д. т. н., професор, 01.05.02

Ідентифікатор ORCID ID: 0000-0001-9318-2859

Додаткова інформація:

Повне найменування юридичної особи: Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"

Код за ЄДРПОУ: 02070921

Місцезнаходження: проспект Берестейський, буд. 37, Київ, 03056, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

Власне Прізвище Ім'я По-батькові:

1. Залевська Ольга Валеріївна

2. Olga V. Zalevska

Кваліфікація: к. т. н., доцент, 05.01.01

Ідентифікатор ORCID ID: 0000-0002-3163-1695

Додаткова інформація:

Повне найменування юридичної особи: Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"

Код за ЄДРПОУ: 02070921

Місцезнаходження: проспект Берестейський, буд. 37, Київ, 03056, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

VIII. Заключні відомості

Власне Прізвище Ім'я По-батькові голови ради	Стіренко Сергій Григорович
Власне Прізвище Ім'я По-батькові головуючого на засіданні	Стіренко Сергій Григорович
Відповідальний за підготовку облікових документів	Мельниченко Артем Васильович
Реєстратор	УкрІНТЕІ
Керівник відділу УкрІНТЕІ, що є відповідальним за реєстрацію наукової діяльності	Юрченко Тетяна Анатоліївна