Облікова картка дисертації

I. Загальні відомості

Державний обліковий номер: 0824U003335

Особливі позначки: відкрита

Дата реєстрації: 06-11-2024

Статус: Наказ про видачу диплома

Реквізити наказу МОН / наказу закладу: Наказ ХНУ імені В. Н. Каразіна № 0302-3к/1703 від 23.12.2024 р.

II. Відомості про здобувача

Власне Прізвище Ім'я По-батькові:

1. Дейнега Олександр Андрійович
2. Oleksandr Deineha

Кваліфікація:

Ідентифікатор ORCID ID: 0000-0001-8024-8812

Вид дисертації: доктор філософії

Аспірантура/Докторантура: так

Шифр наукової спеціальності: 122

Назва наукової спеціальності: Комп’ютерні науки

Галузь / галузі знань: інформаційні технології

Освітньо-наукова програма зі спеціальності: Комп'ютерні науки

Дата захисту: 04-12-2024

Спеціальність за освітою: Комп'ютерні науки

Місце роботи здобувача: Харківський національний університет імені В. Н. Каразіна

Код за ЄДРПОУ: 02071205

Місцезнаходження: майдан Свободи, буд. 4, Харків, Харківський р-н., 61022, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

III. Відомості про організацію, де відбувся захист

Шифр спеціалізованої вченої ради (разової спеціалізованої вченої ради): PhD 7047

Повне найменування юридичної особи: Харківський національний університет імені В. Н. Каразіна

Код за ЄДРПОУ: 02071205

Місцезнаходження: майдан Свободи, буд. 4, Харків, Харківський р-н., 61022, 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: Харківський національний університет імені В. Н. Каразіна

Код за ЄДРПОУ: 02071205

Місцезнаходження: майдан Свободи, буд. 4, Харків, Харківський р-н., 61022, 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, 50.05.09, 20.54.03, 20.54.06

Тема дисертації:

1. Оптимізація функціональних мов програмування на основі методів штучного інтелекту

2. Optimization of Functional Programming Languages Based on Artificial Intelligence Methods

Реферат:

1. Дисертація присвячена оптимізації функціональних мов програмування, на основі методів штучного інтелекту, що є складною та важливою задачею з багатьма проблемами та викликами. В дисертації розглянуто лямбда-числення як приклад відносно простої репрезентації функціональних мов програмування, що дозволяє показати процеси компіляції та інтерпретації функціональних мов програмування шляхом редукції лямбда-термів. У першому розділі описано теоретичну частину дослідження. Описано переваги функціональних програм, такі як простота тестування та надійність коду, а також їх недоліки, основним з яких є низька продуктивність. Пояснюється можливість переходу від роботи з функціональними мовами програмування до лямбда-числення. Далі представлені підходи для оптимізації лямбда-числення, основним із яких є удосконалення стратегій редукції лямбда-термів. Далі текст заглиблюється в зв'язок між лямбда-численням і верифікацією програм в контексті паралельного програмування. Описано важливість формальної перевірки для паралельних програм, особливо з огляду на потенційні складності та проблеми, пов'язані з одночасним виконанням. У другому розділі представлений

підхід до оптимізації стратегій редукції, що базується на змішуванні стратегій та використанні рандомізованих стратегій. Описані результати, що показують ефективність даного підходу, та можливість заміни чистих стратегій змішаними, що дозволяють зберегти існуючу продуктивність, проте підвищити загальну вірогідність успішного редукування термів. Далі у розділі була розглянута концепція обчислювальної нерівнозначності редексів лямбда-термів, що є ключовими точками у виборі стратегії редукції. Нерівнозначність була оцінена з використанням методів машинного навчання для вирішення задачі регресії. Ціллю регресії була оцінка часу виконання операції редукції для даного редексу по параметрам терму, що відображають його деревну структуру. В результаті було отримано відхилення від очікуваного логарифму часу в 0.28 для регресійної моделі на базі штучної нейронної мережі та в 0.28 для лінійної регресії. У третьому розділі була перевірена можливість оцінки кількості кроків редукції лямбда-термів за заданою стратегією із застосуванням методів глибинного навчання. Аналіз проводився з використанням методів глибинного навчання для аналізу послідовностей. Показано, що точних результатів оцінки можливо досягти при визначенні 0-2 кроків редукції проте із збільшенням очікуваних кроків редукції зростала помилка в оцінці. Далі було досліджено можливість використання вбудовувань для репрезентації різниці в редукції лямбда термів різними стратегіями. Для цього було розглянуто чотири моделі LLM для генерації вбудовувань з текстових представлень лямбда-термів. Згенеровані вбудовування були використані для створення восьми наборів даних для кожної розглянутої моделі LLM та для стратегій редукції термів LO та RI. Для оцінки якості репрезентації інформації у вбудовуваннях були використані моделі ШНМ, що вирішували проблему класифікації відносно кроків редукції від 0 до 30. Далі навчені моделі ШНМ були оцінені з показниками для оцінки точності регресії MAE, RMSE. Результати вказують на те, що код і загальні LLM можуть допомогти отримати інформацію з лямбда-термів і використовувати цю інформацію для вибору оптимальної стратегії редукції. У четвертому розділі лямбда-терми були перетворені в усереднені вектори вбудовувань розміром 768, що були отримані в результаті застосування попередньонавченої моделі ШНМ для задач пов'язаних із аналізом програмного коду Microsoft CodeBERT та подальшої обробки виходів середніх рівнів цієї моделі за принципом об'єднання слів у значущі вектори Word2Vec. Далі було досліджене формування кластерів даних із застосуванням методу DBSCAN, що використовує як евклідову, так і косинусну метрику, окрім методу агломеративної кластеризації з використанням евклідової, косинусної, L1 і L2 метрики. Ці зусилля з кластеризації підкреслили ефективність моделі CodeBERT у вилученні значущих характеристик із лямбда-термів. Також було продовжено ідею трансформації лямбда-термів у вектори вбудовувань з використанням моделей OpenAI з розміром векторів 1536, та 3072. Дані вектори були так само проаналізовані з застосуванням методів PCA та t-SNE для візуалізації цих векторів. В наступній частині четвертого розділу представлено підхід для використання LLM безпосередньо для проведення процесу редукції лямбда-термів. Результати показали, що використання LLM для вирішення цієї задачі не є достатньо ефективним. Далі представлено можливий варіант імплементації описаних методів для використання у компіляторах для підвищення їх продуктивності. Сукупність результатів, викладених у дисертації, разом із підтвердженою науковою та практичною актуальністю демонструють досягнення поставленої мети щодо оптимізації функціональних мов програмування на базі методів штучного інтелекту.

2. The dissertation is devoted to the optimization of functional programming languages based on artificial intelligence methods, which is a complex and important task with many problems and challenges. The thesis examines lambda calculus as an example of a relatively simple representation of functional programming languages, which allows us to show the processes of compilation and interpretation of functional programming languages by reducing lambda terms. The first chapter describes the theoretical part of the research. The advantages of functional programs, such as ease of testing and code reliability, are described, as well as their disadvantages, the main of which is low performance. The possibility of transition from working with functional programming languages to lambda calculus is explained. Next, approaches for optimizing the lambda calculus are presented, the main of which is the improvement of strategies for the reduction of lambda terms. Later in this section, the text delves into the connection between lambda calculus and program verification in the context of parallel programming. This includes verifying properties such as safety, liveness, and correctness in various

execution scenarios. The second chapter presents an approach to optimization of reduction strategies based on mixing strategies and using randomized strategies. The results are described, showing the effectiveness of this approach and the possibility of replacing pure strategies with mixed ones, which allow maintaining the existing performance, but increasing the overall probability of successful term reduction. Further in the section, the concept of computational inequality of lambda-term redexes, which are key points in choosing a reduction strategy, was considered. The disparity was estimated using machine learning techniques to solve the regression problem. The purpose of the regression was to estimate the time of the reduction operation for a given redex based on the parameters of the term reflecting its tree structure. In the third chapter, the possibility of estimating the number of lambda term reduction steps according to a given strategy using deep learning methods was tested. The analysis was carried out using deep learning methods for sequence analysis. It is shown that accurate estimation results can be achieved when determining 0–2 reduction steps, however, with the increase of the expected reduction steps, the error in the estimation increased. Next, the possibility of using embeddings to represent the difference in the reduction of lambda terms by different strategies was investigated. For this, four LLM models were considered for generating embeddings from text representations of lambda terms. The generated embeddings were used to generate eight datasets for each LLM model considered and for the LO and RI term reduction strategies. ANN models were used to assess the quality of information representation in embeddings, which solved the classification problem with respect to reduction steps from 0 to 30. Next, the trained ANN models were evaluated with indicators for assessing the accuracy of regression MAE, RMSE. Such results indicate that the code and general LLMs can help extract information from lambda terms and use this information to select an optimal reduction strategy. In the fourth chapter, lambda terms were transformed into averaged embedding vectors of size 768, which were obtained as a result of applying a pre-trained ANN model for tasks related to the analysis of Microsoft CodeBERT software code and further processing of the outputs of the average levels of this model according to the principle of combining words in significant Word2Vec vectors. Therefore, the formation of data clusters using the DBSCAN method using both Euclidean and cosine metrics, in addition to the agglomerative clustering method using Euclidean, cosine, L1 and L2 metrics, was further investigated. This clustering effort highlighted the effectiveness of the CodeBERT model in extracting meaningful features from lambda terms. The idea of transforming lambda terms into embedding vectors was also continued using OpenAI models with vector sizes of 1536 and 3072. These vectors were also analyzed using PCA and t-SNE methods to visualize these vectors. The next part of the fourth chapter presents an approach for using the LLM directly to perform the lambda term reduction process. The last chapter of the dissertation presents a possible implementation option. The set of results presented in the dissertation, together with the confirmed scientific and practical relevance, demonstrate the achievement of the set goal of optimizing functional programming languages based on artificial intelligence methods.

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:

- Deineha, O., Donets, V., & Zholtkevych, G. (2024). The approach development of data extraction from lambda terms. Eastern-European Journal of Enterprise Technologies.
- Deineha, O. (2024). Supervised data extraction from transformer representation of Lambda-terms. Radioelectronic and Computer Systems, 2024(2), 19-29.
- Deineha O. The Clustering of Lambda Terms by Using Embeddings. Вісник Харківського національного університету імені В.Н. Каразіна, сер. «Математичне моделювання. Інформаційні технології.

Автоматизовані системи управління». 2023. вип. 59. С.16-23.

- Deineha, O. (2024). Lambda calculus term reduction: Evaluating LLMS' predictive capabilities. Information Technology and Society, 1(12), 51-55.

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

Соціально-економічна спрямованість:

Охоронні документи на ОПІВ:

Впровадження результатів дисертації: Впроваджено

Зв'язок з науковими темами: 0121U109183

VI. Відомості про наукового керівника/керівників (консультанта)

Власне Прізвище Ім'я По-батькові:

1. Жолткевич Григорій Миколайович

2. Grygoriy Zholtkevych

Кваліфікація: д. т. н., професор, 05.02.08

Ідентифікатор ORCID ID: 0000-0002-7515-2143

Додаткова інформація:

Повне найменування юридичної особи: Харківський національний університет імені В. Н. Каразіна

Код за ЄДРПОУ: 02071205

Місцезнаходження: майдан Свободи, буд. 4, Харків, Харківський р-н., 61022, 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. Шаронова Наталія Валеріївна

2. Nataliia Sharonova

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

Ідентифікатор ORCID ID: 0009-0004-9878-1761

Додаткова інформація:

Повне найменування юридичної особи: Національний технічний університет "Харківський політехнічний інститут"

Код за ЄДРПОУ: 02071180

Місцезнаходження: вул. Кирпичова, буд. 2, Харків, Харківський р-н., 61002, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

Власне Прізвище Ім'я По-батькові:

1. Шаховська Наталія Богданівна

2. Nataliya Shakhovska

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

Ідентифікатор ORCID ID: 0000-0002-6875-8534

Додаткова інформація:

Повне найменування юридичної особи: Національний університет "Львівська політехніка"

Код за ЄДРПОУ: 02071010

Місцезнаходження: вул. Степана Бандери, буд. 12, Львів, 79013, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

Рецензенти

Власне Прізвище Ім'я По-батькові:

1. Узлов Дмитро Юрійович

2. Dmytro Uzlov

Кваліфікація: к. т. н., 05.13.06

Ідентифікатор ORCID ID: 0000-0003-3308-424X

Додаткова інформація:

Повне найменування юридичної особи: Харківський національний університет імені В. Н. Каразіна

Код за ЄДРПОУ: 02071205

Місцезнаходження: майдан Свободи, буд. 4, Харків, Харківський р-н., 61022, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

Власне Прізвище Ім'я По-батькові:

1. Меняйлов Євген Сергійович

2. Meniailov Yevhen S.

Кваліфікація: к. т. н., 01.05.02

Ідентифікатор ORCID ID: Не застосовується

Додаткова інформація:

Повне найменування юридичної особи: Харківський національний університет імені В. Н. Каразіна

Код за ЄДРПОУ: 02071205

Місцезнаходження: майдан Свободи, буд. 4, Харків, Харківський р-н., 61022, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

VIII. Заключні відомості

Власне Прізвище Ім'я По-батькові голови ради	Толстолузька Олена Геннадіївна
Власне Прізвище Ім'я По-батькові головуючого на засіданні	Толстолузька Олена Геннадіївна
Відповідальний за підготовку облікових документів	Шевченко Андрій Олександрович
Реєстратор	УкрІНТЕІ
Керівник відділу УкрІНТЕІ, що є відповідальним за реєстрацію наукової діяльності	Юрченко Тетяна Анатоліївна