МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования
«Дальневосточный федеральный университет»
(ДВФУ)
Институт математики и компьютерных технологий (Школа)
Академия цифровой трансформации

Петров Сергей Дмитриевич

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
Магистерская диссертация

ИЗВЛЕЧЕНИЕ ПРИЗНАКОВОГО ПРЕДСТАВЛЕНИЯ ИСХОДНОГО КОДА С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ ОБУЧЕНИЯ БЕЗ УЧИТЕЛЯ ДЛЯ DOWNSTREAM ОБУЧЕНИЯ МОДЕЛЕЙ

по направлению подготовки
09.04.01 «Информатика и вычислительная техника»,
магистерская программа «Искусственный интеллект и большие данные»

Владивосток
2025

В материалах данной выпускной квалификационной работы не содержатся сведения, составляющие государственную тайну, и сведения, подлежащие экспортному контролю

Автор работы _____
(подпись)
Группа М9119-09.04.01иибд
«_07_»_____июля_____ 2021 г.

Уполномоченый по экспортному контролю

_____ Е. В. Сапрыкина
(подпись) (И.О.Ф.)
«_07_»_____июля_____ 2021 г.

Руководитель ВКР _____
(должность, уч. степень, уч. звание)
_____ А. Г. Тыщено
(подпись) (И.О.Ф.)
«_07_»_____июля_____ 2021 г.

Консультант

_____ А. Г. Тыщено
(подпись) (И.О.Ф.)
«_07_»_____июля_____ 2021 г.

Назначен рецензент
зав. лаб. НЦВИ ИОФ РАН, к.ф.-м.н.
(уч. степень, уч. звание)
Луньков Андрей Аллександрович
(фамилия, имя, отчество)

Защищена в ГЭК с оценкой

_____

Секретарь ГЭК

_____ Т. С. Тихонова
(подпись) (И.О.Ф.)
«_07_»_____июля_____ 2021 г.

«Допустить к защите»

Академии цифровой трансформации, к.э.н.

_____ Е. В. Сапрыкина
(подпись) (И.О.Ф.)
«_07_»_____июля_____ 2021 г.

Аннотация

Данная выпускная квалификационная работа посвящена исследованию методов обучения без учителя для извлечения признаковых представлений исходного кода с целью их дальнейшего использования в downstream-задачах машинного обучения. В современных условиях разработки программного обеспечения анализ и обработка исходного кода играют ключевую роль в таких задачах, как предсказание дефектов, автоматический рефакторинг, классификация кода и поиск уязвимостей. Однако эффективное представление кода в машинно-читаемом формате остается сложной задачей, требующей применения современных методов искусственного интеллекта.

Цель данной работы заключается в адаптации алгоритма самообучения DINO для работы с текстовыми данными, в частности с исходным кодом, и сравнительном анализе его эффективности с готовыми моделями представления кода. В рамках исследования был проведен анализ алгоритма, предложена его модификация для обработки текстовых последовательностей, обучены векторные представления исходного кода и выполнена их оценка на downstream-задачах, включая классификацию кода.

Результаты работы демонстрируют потенциал методов обучения без учителя для автоматического извлечения информативных признаков из исходного кода. Разработанные подходы могут быть интегрированы в инструменты статического анализа, системы контроля качества кода и другие решения, направленные на повышение эффективности разработки программного обеспечения.

СОДЕРЖАНИЕ

Введение

В современной разработке программного обеспечения исходный код является ключевым ресурсом, требующим эффективного анализа и обработки. С ростом сложности программных систем и увеличением объёмов кодовой базы традиционные методы анализа кода сталкиваются с рядом ограничений, связанных с масштабируемостью и точностью. В таких задачах, как автоматическое обнаружение уязвимостей, рефакторинг, поиск семантически схожих фрагментов кода и предсказание дефектов, критически важным становится наличие качественного признакового представления исходного кода, которое могло бы быть использовано в downstream-моделях машинного обучения.

Актуальность задачи

1. TODO Актуальность задачи

Цель работы

Адаптация алгоритма DINO для извлечения признаковых представлений исходного кода и сравнительный анализ его эффективности с готовыми моделями (CodeBERT) на downstream-задачах, таких как классификация кода.

Задачи исследования

1. Провести обзор современных методов представления исходного кода и алгоритмов самообучения (self-supervised learning).
2. Модифицировать алгоритм DINO для работы с текстовыми данными.
3. Собрать и предобработать датасеты для обучения и оценки моделей.
4. Обучить модель на основе адаптированного DINO и сравнить её с существующими решениями и проанализировать результаты.

Структура работы

1 Описание предметной области

1.1 Машинное обучение

Машинное обучение (ML) – это раздел искусственного интеллекта, изучающий методы построения алгоритмов, способных автоматически обучаться и улучшать свою работу на основе данных без явного программирования. В отличие от традиционных алгоритмов, где поведение системы жестко задаётся разработчиком, модели машинного обучения выявляют закономерности в данных и используют их для прогнозирования, классификации или принятия решений

Результат обучения алгоритма называется моделью – параметризированное отражение(функция), которое преобразует объекты из пространства входных признаков в пространство предсказаний.

Одним из главных требований к модели – ее способность к обобщению. Благодаря этому модель не просто запомнит данные на которых училась, а находит в них закономерности, что позволяет более точно делать отражение на новых объектах.

Алгоритмы ML делятся на три основные категории:

– Обучение с учителем

– Обучение с учителем

– Обучение с подкреплением

1.1.1 Обучение с учителем

Обучение с учителем – это вид машинного обучения, при котором модель обучается на примерах, где для каждого входного объекта известен правильный ответ. Цель такого метода обучения – построение модели которая будет способна предсказать ответ для ранее не встречавшихся примеров с заданой точностью.

Основное преимущество обучения с учителем:

– **Интерпретируемость** – модель работает с зарание определенным пространством выходных значений, так как разметка чаще составляется

человеком.

– **Интуитивная оценка качества** – для оценки часто используется интуитивно понятные метрики такие как точность или средний квадрат ошибки.

Основные недостатки обучения с учителем:

– **Зависимость от качества входных данных** – эффективность модели напрямую определяется качеством размеченных данных, процесс создания которых требует значительных временных и трудозатрат. Для сложных задач объём требуемых данных может возрастать экспоненциально, что создаёт существенные практические ограничения.

– **Проблема переобучения** – существует риск избыточной подгонки модели под особенности обучающей выборки – в таком случае алгоритм начинает воспроизводить не только значимые закономерности, но и случайные шумы, что резко снижает его способность к обобщению на новых данных.

Пример модлей которые обучаются с помощью обучения с учителем:

– Линейная регрессия

– Дерево решений

– Метод опорных векторов

1.1.2 Обучение без учителя

Обучение без учителя – это вид машинного обучения, при котором модель обучается на примерах для которых нет какой-либо разметки. Цель такого метода обучения – построение модели которая сама будет находить закономерности, не опираясь на внешние подсказки.

Задачи обучения без учителя включают в себя:

– **Кластеризацию** – задача разделения объектов на группы, которые имеют сходство между собой и отличаются от других.

– **Снижение размерности** – задача уменьшение количества признаков данных, сохраняя при этом информацию об объекте.

– **Поиск ассоциативных правил** – задача выявления устойчивых взаимосвязей между событиями в больших данных.

– **Генеративные модели** – задача генерации новых данных похожих на тренировочные.

Основное преимущество обучения без учителем:

– **Не требуется разметка данных** – работа с неразмеченными данными значительно облегчает процесс сбора данных.

– **Гибкость и универсальность** – применимо в разнообразных областях, а так же может использоваться для предобработки данных перед обучением с учителем

Основные недостатки обучения без учителем:

– **Сложность оценки качества** – отсутсвие разметки затрудняет объективную оценку результатов.

– **Проблема интерпретируемости** – так как пространство выходных значений не известно, сложно их интерпретировать.

Пример модлей которые обучаются с помощью обучения с учителем:

– К средних

– Генеративные состязательные сети

1.1.3 Обучение с подкреплением

Обучение с подкреплением – это вид машинного обучения, при котором агент(модель) обучается на основе опыта взаимодействия со средой, принимая решения которые максимизируют награду.

В отличии от прошлых методов, агенты ориентированы на последовательное принятие решений в условиях неопредленности.

Основное преимущество обучения с подкреплением:

– **Подходит для задач с отложенной наградой** – Может учитывать долгосрочные последствия действий, а не только мгновенную выгоду.

– **Возможность обучения без размеченных данных** – Не требует готовых ”правильных ответов”.

Основные недостатки обучения с подкреплением:

– **Высокие вычислительные затраты** – Требует миллионов (иногда миллиардов) попыток для обучения.

– **Проблема исследования-эксплуатации** – Агент должен балансировать между исследованием и эксплуатацией. Исследование – проба новых действий, для поиска лучшей стратегии. Эксплуатация – использование уже известных лучших действий.

Пример модлей которые обучаются с помощью обучения с учителем:

– К средних

– Генеративные состязательные сети