БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Факультет прикладной математики и информатики

Кафедра дискретной математики и алгоритмики

Аннотация к дипломной работе

«Метод обнаружения преднамеренно искаженных данных в системах машинного обучения»

Лозинская Ирина Михайловна

Научный руководитель – старший преподаватель кафедры дискретной математики и алгоритмики ФПМИ Комаровский И. В.

Минск, 2025

АННОТАЦИЯ

Дипломная работа, 50 страниц, 3 таблицы, 16 иллюстраций, 18 формул, 32 источника.

Ключевые слова: ТРИГГЕР, МАСКА, НЕЙРОННЫЙ СЛОЙ, МАШИННОЕ ОБУЧЕНИЕ, ОТРАВЛЕНИЕ ДАННЫХ, МЕТОДЫ ЗАЩИТЫ.

Объект исследования является проблема определения наличия отравления в не доверенном наборе данных для тренировки моделей машинного обучения.

Предметом исследования являются методы и алгоритмы обнаружения отравленных изображений и идентификации триггеров в составах датасетов, используемых для обучения моделей машинного обучения.

Целью работы является разработка и реализация алгоритма определения отравленных изображений и выделения из них триггера.

Методами исследования являются методы машинного обучения и методы оптимизации.

Полученные результаты и их новизна: выявлены ограничения метода Activation Clustering при моделировании атак, заключающихся в использовании единого триггера для перенаправления всех исходных классов в целевой класс. Для повышения точности идентификации отравленных данных был разработан и реализован дополнительный метод с применением модели CatBoost. Кроме того, в ходе оценки эффективности Neural Cleanse обнаружены существенные недостатки в восстановлении триггеров, охватывающих значительную часть изображения. В ответ на эти ограничения был предложен и реализован усовершенствованный подход, обеспечивающий более качественное восстановление триггеров большой площади.

Достоверность материалов и результатов дипломной работы: использованные материалы и результаты дипломной работы являются достоверными. Работа выполнена самостоятельно.

Областью возможного практического применения является защита процесса тренировки моделей машинного обучения от отравленных данных.

АНАТАЦЫЯ

Дыпломная праца, 50 старонак, 3 табліцы, 16 ілюстрацый, 18 формул, 32 крыніцы.

Ключавыя словы: ТРУГЕР, МАСКА, НЕЙРОНАВЫ ПЛАСТ, МАШЫННАЕ НАВУЧАННЕ, АТРУЧВАННЕ ДАДЗЕНЫХ, МЕТАДЫ АБАРОНЫ.

Аб'ектам даследавання з'яўляецца праблема вызначэння наяўнасці атручвання ў не даверанам наборы дадзеных для трэніроўкі мадэляў машыннага навучання.

Прадметам даследавання з'яўляюцца метады і алгарытмы выяўлення атручаных малюнкаў і ідэнтыфікацыі трыгераў ў складах датасетов, якія выкарыстоўваюцца для навучання мадэляў машыннага навучання.

Мэтай даследвання з'яўляецца распрацоўка і рэалізацыя алгарытму вызначэння атручаных малюнкаў і вылучэнні з іх трыгера.

Метадамі даследавання з'яўляюцца метады машыннага навучання і метады аптымізацыі.

Атрыманыя вынікі і іх навізна: выяўлены абмежаванні метаду Activation Clustering пры мадэляванні нападаў, якія складаюцца ў выкарыстанні адзінага трыгера для перанакіравання ўсіх зыходных класаў у мэтавы клас. Для павышэння дакладнасці ідэнтыфікацыі атручаных дадзеных быў распрацаваны і рэалізаваны дадатковы метад з ужываннем мадэлі CatBoost. Акрамя таго, у ходзе ацэнкі эфектыўнасці Neural Cleanse выяўлены істотныя недахопы ў аднаўленні трыгераў, якія ахопліваюць значную частку малюнка. У адказ на гэтыя абмежаванні быў прапанаваны і рэалізаваны удасканалены падыход, які забяспечвае больш якаснае аднаўленне трыгераў вялікай плошчы.

Даставернасць матэрыялаў і вынікаў дыпломнай працы: выкарыстаныя матэрыялы і вынікі дыпломнай Працы з’яўляюцца даставернымі. Праца выканана самастойна.

Вобласцью магчымага практычнага прымянення з'яўляецца абарона працэсу трэніроўкі мадэляў машыннага навучання ад атручаных дадзеных.

ANNOTATION

Diploma work, 50 pages, 3 tables,16 illustrations, 18 formulas, 32 sources.

Keywords: TRIGGER, MASK, NEURAL LAYER, MACHINE LEARNING, DATA POISONING, PROTECTION METHODS.

The object of the research is the problem of determining the presence of poisoning in an untrusted dataset for training machine learning models.

The subject of the research is methods and algorithms for detecting poisoned images and identifying triggers in dataset compositions used to train machine learning models.

The purpose of the research is development and implementation of the algorithm for detecting poisoned images and isolating a trigger from them.

Methods of research are machine learning methods and optimization methods.

The results of the work and their novelty: the limitations of the Activation Clustering method in attack modeling have been identified, which consist in using a single trigger to redirect all source classes to the target class. To improve the accuracy of identification of poisoned data, an additional method was developed and implemented using the CatBoost model. In addition, during the evaluation of the effectiveness of Neural Cleanse, significant shortcomings were found in the restoration of triggers covering a significant part of the image. In response to these limitations, an improved approach has been proposed and implemented to provide better recovery of large-area triggers.

Authenticity of the materials and results of the diploma work: the materials used and the results of the diploma work are authentic. The work has been put through independently.

Recommendations on the usage. The results of the work can be used to protect the process of training machine learning models from poisoned data.