Министерство образования Республики Беларусь

Учреждение образования

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

УДК 004.891.3

ПУНЬКО
Владислав Валерьевич

Система автоматического анализа биопсий

АВТОРЕФЕРАТ

на соискание степенни магистра по
специальности 1-40 80 04 – Информатика и технология программирования

Научный руководитель
доцент, кандидат технических наук
Волорова Н. А.

Минск 2021

КРАТКОЕ ВВЕДЕНИЕ

Сегодня уже трудно представить деятельность человека в каком-то определенном направлении без использования вычислительной техники. Огромный скачок мощностей компьютеров за последние десять лет, а также доступность и уменьшение их габаритов позволило применять вычислительную технику практически где угодно. Это легко доказать путем сравнения числа применяемой техники за последние годы в медицине, строительстве, сельском хозяйстве и в других областях, не связанных напрямую с ними.

Без всяких сомнений, такое сильное развитие технологий, связанное с вычислительной техникой, напрямую повлияло на изменения программного обеспечения. В свою очередь, это послужило поводом для роста интереса к интеллектуальным системам, основанным на принципах искусственных нейронных сетей. С момента создания первых таких систем, нейронные сети стали использоваться как для решения различных прикладных задач, так и для проведения исследований. Данная технология успешно зарекомендовала себя в широком спектре задач, таких как распознавание образов, прогнозирование и классификация, сжатие и извлечение информации и другие.

В последние несколько лет наблюдается прирост пациентов с различными онкологическими заболеваемости. На этот вывод могут влиять сразу несколько факторов: ухудшение глобального климата, загрязнение, а также большее количество правильно диагностируемых заболеваний, что значительно повлияло на общую статистику. Несомненно, в настоящее время разработаны новые подходы в диагностике и лечении таких заболеваний на ранней стадии. Некоторые типы рака, ранее неизлечимые, сейчас можно отнести к излечимым, однако большинство типов таких заболеваний не могут быть идентифицированы на ранних стадиях, из-за чего спасти пациента невозможно.

Исследования, нацеленные на прогнозирование и выявление раковых заболеваний на ранних стадиях проводятся уже более 20 лет. За это время были предложены многие методы и подходы для решения данной проблемы путем полного анализа генов пациента, анализ изображений биопсий и других материалов. Были реализованы программы помощи врачам в анализе, ведь от корректной стратегии лечения зависит не только его эффективность, но и жизнь человека.

Стоит отметить, что в медицине существует большое количество задач, где человек не может показывать удовлетворительный результат, а в нередких случаях вообще не способен решить поставленную задачу. Это связано в первую очередь с тем, что человеческий мозг легко обмануть. Помимо этого, долгая и кропотливая работа может выводить сенсоры из нормального состояния, что напрямую отразится на работе (к примеру появление боли в глазах после длительной работы с мелкими объектами).

Спектр задач, решаемых с применением искусственного интеллекта в медицине, значительно вырос и, помимо предсказания или классификации заболеваний, уже есть решенные задачи, связанные с распознаванием меланомы, сегментацией и выделением пораженных участков легких и прочее.

В настоящее время имеется как минимум шесть хорошо изученных типов рака, которые могут быть обработаны с лучшей точностью, чем эксперт (меланома, рак молочной железы), что дает предпосылки для проведения различных исследований и осуществления попыток «победить» тот или иной тип рака. Таким образом, основной задачей данного проекта является попытка спроектировать инструмент, дешевый в производстве и способный решать определенный спектр задач, шире, чем есть на данный момент (с минимальными изменениями).

Вследствие вышесказанного, можно сказать, что объектом исследования являются изображения биопсий и методы их обработки, существующие на данный момент. Предметом в свою очередь служит разработка комплекса для упрощения анализа подобных изображений.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Постановка задач исследования

Целью данного исследовательского проекта является разработка алгоритма для автоматизации процесса распознавания раковых клеток, а также реализация его программно-аппаратной части, которая в дальнейшем будет использоваться в медицинских диагностических центрах. Для достижения поставленной цели необходимо выполнить следующие условия:

- алгоритм должен подстраиваться под разнообразные виды данных;
- алгоритм подготовки данных должен быстро выполнять функционал;
- алгоритм сегментации не должен использовать много памяти;
- классификатор должен иметь подстраховку на случай ошибки;
- работа с продуктом должна быть ассоциативно понятной;
- добавить поддержку нескольких заболеваний;
- переоборудовать световой микроскоп в электронный;
- реализовать статистический анализ здоровых и больных клеток;
- реализовать функционал формирования отчетности.

Таким образом, поставленная задача разбивается на два логических и функциональных блока: создание аппаратного и программного модуля. Для реализации программного модуля проекта сперва необходимо изучить существующие функциональные аналоги для упрощения и ускорения разработки. После чего, на основании произведенного анализа, необходимо выбрать и изучить технологии для реализации программной части исследовательского проекта. Далее спроектировать гибкое архитектурное решение для алгоритма подготовки исходных данных, сегментации изображения, классификации раковых клеток по заболеваниям. Затем реализовать спроектированные алгоритмы, протестировав их на тестовых данных, рассчитать аппаратную часть проекта, составить схему питания и создать трехмерные модели деталей для печати. И напоследок реализовать аппаратную часть, собрать цепь питания для микрокомпьютера, интегрировать дополнительные детали для подсветки микропрепарата и откалибровать её под реализованное программное обеспечение.

Обзор существующих аналогов проекта

Существует несколько решений для автоматизации процесса диагностики раковых заболеваний с использованием искусственных нейронных сетей. Ниже приведены некоторые из задач, которые могут выполнять рассматриваемые аналоги программного продукта:

- диагностировать не только раковые заболевания;
- подготавливать отчетность для врачей-специалистов;

- предсказывать дальнейшее развитие болезни;
- применять различные алгоритмы статистического анализа клеток;
- проводить диагностику нескольких онкологических заболеваний.

Наибольший интерес для исследования в существующем программном обеспечении представляет функциональность применения статистического анализа клеток. На сегодняшний момент на рынке существует два аналога разрабатываемого продукта. Рассмотрим каждый из них по отдельности.

Watson является когнитивной компьютерной системой, разработанной в компании IBM, которая обрабатывает информацию скорее как человек, чем компьютер. Понимая естественный язык, она генерирует множество гипотез, основанных на фактических данных, и изучает их в процессе решения поставленной задачи. Такая система когнитивных вычислений была специально разработана для поддержки сообщества онкологов и повышения эффективности их работы. С помощью этой системы медицинские эксперты могут использовать свой опыт исследования в диагностике разнообразных заболеваний. Также они могут интерпретировать клиническую информацию пациентов и назначать индивидуальные варианты лечения, основанные на фактических данных.

AR-микроскоп представляет собой улучшенный световой микроскоп с функцией дополненной реальности, который был разработан в компании Google с целью облегчить диагностику онкологических заболеваний. Этот аппаратный комплекс способен обнаруживать раковые клетки в режиме реального времени, используя машинное обучение.

В аппаратном модуле используется модифицированный световой микроскоп, интегрированный с алгоритмами анализа изображений и машинного обучения. Дисплей дополненной реальности находится над камерой, которая связывается с алгоритмом для отображения данных, как только обнаруживает проблему. Другими словами, микроскоп сразу обнаруживает раковые клетки, как только вы помещаете образец под его объективом. Он эффективно выполняет ту же работу, что и врач, но намного быстрее.

В будущем данная разработка станет легкодоступной, а также будет интегрироваться в существующие устройства, не требуя доработки. В конечном итоге использование аппаратного модуля будет возможно для диагностики и других заболеваний, таких как туберкулез и малярия. К сожалению, рассматриваемый проект находится в разработке и о конечном завершении в перспективе пока ничего не известно.

Личный вклад соискателя

Результаты, приведенные в диссертационной работе, и положения, выносимые на защиту, получены соискателем лично. Вклад научного руководителя доцента, кандидата технических наук Натальи Алексеевны Волоровой связан с постановкой целей и задач исследования, определением возможных путей ре-

шения и обсуждением результатов исследований, проводимых автором. В публикациях с соавторами вклад соискателя определяется рамками излагаемых в диссертации результатов.

Апробация результатов диссертации

Основные результаты диссертационной работы докладывались и обсуждались на 2 международных и республиканских научных конференциях: «МЕДЭЛЕКТРОНИКА-2018» — Минск, Республика Беларусь, 2018; «14th International Conference on Pattern Recognition and Information Processing» — Минск, Республика Беларусь, 2019.

Структура и объем диссертации

Диссертационная работа состоит из введения, общей характеристики работы, трех глав, заключения, библиографического списка и двух приложений. Общий объем диссертационной работы составляет 72 страницы, из них 53 страниц основного текста, 28 рисунка на 23 страницах, библиография из 34 наименований, включая 3 публикации автора и три приложения на 6 страницах.

ОСНОВНОЕ СОДЕРЖАНИЕ

Во **введении** диссертации определена область и указаны основные направления проведенного исследования, показана актуальность выбранной темы диссертационной работы, дана краткая характеристика исследуемых вопросов, обозначена практическая ценность работы.

В **первой главе** проводится подробный анализ предметной области задачи, поставленной в проекте, рассматриваются вопросы сущности искусственных нейронных сетей и принципов их работы. Производится оценка сложности проблем, образующихся при использовании искусственных нейронных сетей для решения различных прикладных задач. Также рассматриваются принципы работы реализованных в программных модулях алгоритмов распознавания и сегментации образов на изображениях и производится сравнительный анализ их эффективности для решения подобных задач.

Нейронные сети, или искусственные нейронные сети, являются технологией, объединяющей сразу несколько дисциплин: линейную алгебру, математическую статистику, теорию вероятностей, а также нейрофизиологию и различные отделы компьютерных наук. Они находят широкое применение в различных областях науки, таких как математическое моделирование, анализ больших объемов данных, распознавание и классификация образов, обработка разнообразных сигналов, предсказание и анализ временных рядов. Данная распространенность искусственных нейронных сетей обусловлена их самым главным свойством — способностью обучаться, используя существующие данные, при участии учителя или же без него.

Изучение и исследование искусственных нейронных сетей обусловлены в первую очередь радикальным отличием подходов, используемых биологическим мозгом в обработке получаемых данных, от методов, применяемых в обычной вычислительной технике. Любой биологический мозг, безусловно, можно назвать довольно сложным и нелинейным компьютером, который, в свою очередь, может быстро решать поставленные перед ним задачи. Он имеет возможность реорганизовывать свои структурные составляющие, известны как нейроны, таким образом, чтобы выполнять определенные вычисления намного быстрее и эффективнее, нежели самый быстрый из существующих на сегодняшний день суперкомпьютеров.

Результаты исследований, проведенных в этих направлениях, отражены в работах Buduma Nikhil, Lacascio Nicholas, Хайкин Саймон, Goodfellow Ian, Bengio Yoshua, Courville Aaron, Patterson Josh, Gibson Adam, Glorot Xavier, Boureau Y-Lan, Ponce Jean, LeCun Yann и др.

Вторая глава посвящена подбору используемых технологий для решений поставленной задачи так как это является неотъемлемым и важным этапом в разработке любого программного продукта. От платформы и языка программирования, на котором будет реализовано некоторое программное

обеспечение, зависит качество кода и производительность труда программистов. Большинство разработчиков программного обеспечения выбирают технологии по субъективным причинам, не понимая сложности проекта, сроков, а иногда и не имея опыта работы с выбранными принципами разработки. Зачастую правильно подобранные инструменты разработки могут сэкономить довольно много времени, сил и финансовых затрат на разработку.

Существует множество факторов, которые могут повлиять на выбор технологий для разработки программного продукта. К примеру, разрабатываемый программный модуль должен работать на операционных системах семейства UNIX и иметь совместимость с их новыми версиями. Также разрабатываемый модуль должен быть реализован с использованием современных инструментов для программирования искусственных нейронных сетей, обеспечивающих высокую производительность и надёжность. Одним из факторов может стать тот язык программирования, которым хорошо владеет имеющийся разработчик. Ключевым же фактором является то, что дальнейшую разработку и поддержку продукта, возможно, будут осуществлять другие разработчики, не принимавшие участие в проектировании приложения и выпуске первой версии продукта.

В связи с тем, что имеющееся программное обеспечение разработано только под операционные системы семейства UNIX, разработку модуля целесообразно продолжать на операционной системе Fedora. Приняв во внимание тот факт, что дальнейшей поддержкой этого программного продукта, возможно, будет заниматься другая команда программистов, желательно не использовать малоизвестные и сложные языки программирования. Исходя из всего этого, выбор основного языка программирования для проекта сужается до трех интерпретируемых, популярных на момент разработки и имеющих необходимый инструментарий языков, а именно Ruby, Python и JavaScript.

В разработке данного программного модуля была задействована внешняя библиотека Keras для искусственных нейронных сетей. Язык программирования Ruby не может её использовать, поэтому его можно исключить из рассмотрения. Для программирования системных приложений не используется язык программирования JavaScript, так как он не поддерживает серию стандартов POSIX и уступает в удобстве использования оставшемуся кандидату. Из этого следует, что Python является первостепенным, кроссплатформенным и элегантным языком программирования, предоставляемым вместе почти со всеми UNIX-подобными системами и подходящим для выполнения поставленной задачи в рассматриваемом проекте.

Для решения поставленной задачи необходимо использовать внешние библиотеки. Для реализации сложных тензорных вычислений, производимых в искусственных нейронных сетях, лучше всего использовать внешнюю библиотеку Keras. Данная библиотека находится в открытом доступе и имеет много

дополнительных инструментов, которые значительно ускоряют разработку и развертывание больших интеллектуальных приложений.

Для работы с классическим машинным обучением лучше всего использовать внешнюю библиотеку Scikit-Learn, так как она предоставляет большой функционал для манипуляции и анализа данных. Используя инструментарий этой библиотеки, можно с легкостью использовать алгоритмы машинного обучения в решаемых задачах.

В разработке данного программного модуля задействована обработка изображений, а также потоковая обработка видео. Для выполнения такой работы лучше всего использовать внешнюю библиотеку OpenCV, которая позволяет с легкостью выполнять математические преобразования с разнообразной графикой. Кроме всего прочего, библиотека имеет большой функционал и является открытой для использования в различных проектах.

Третья глава посвящена архитектуре и модулям системы. Проведение анализа начинается с гистологического изображения, на котором находится межклеточное вещество и различные клетки, среди которых необходимо найти больные, если таковые имеются. Задачу гистологического исследования можно разбить на три основных этапа. На первом этапе все клетки отделяются от межклеточного вещества и других объектов. На втором они сортируются на здоровые клетки и больные, что значительно ускоряет работу системы. На третьем отобранные клетки классифицируются по заболеваниям. С учетом перечисленных этапов оптимальная архитектура программного модуля приведена на рисунке 1, где все процессы (классификация и сегментация клеток) происходит одновременно.

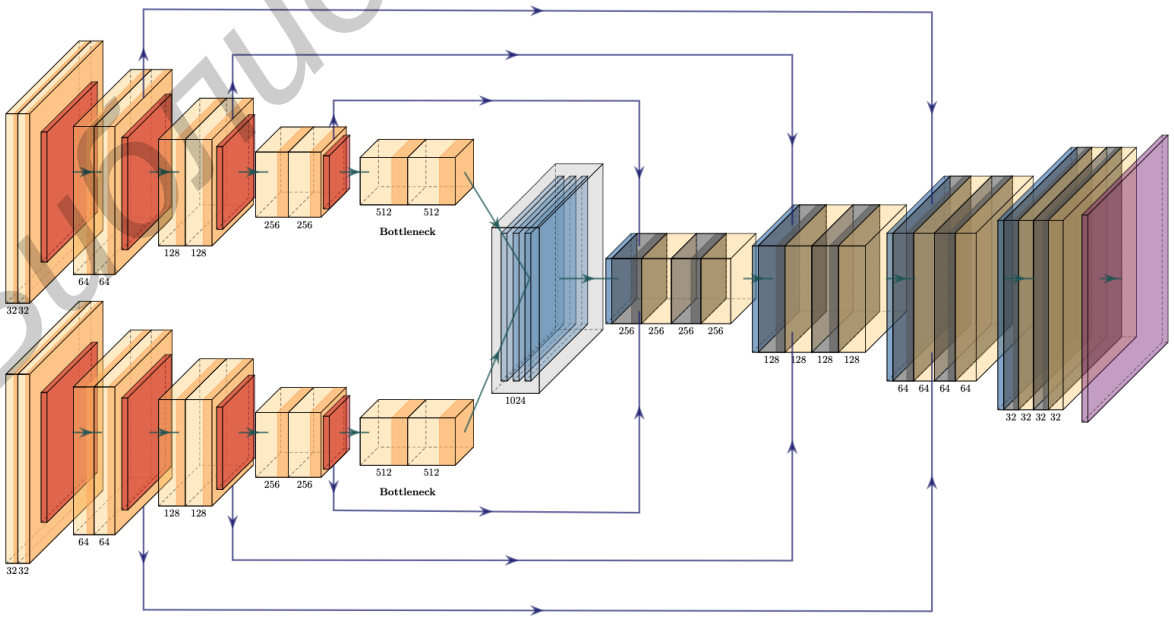

Рисунок 1 — Схема архитектуры разработанного программного модуля

Использование обычных детерминированных алгоритмов будет неэффективным в силу того, что поставленная в проекте задача является довольно специфической и для принятия решения следует учитывать множество факторов. Каждая новая биопсия является уникальной, поэтому для решения поставленной задачи необходимы интеллектуальные алгоритмы, которые могут давать ответ, опираясь на опыт предыдущих результатов. Данный факт повлиял на выбор использования искусственных нейронных сетей в рассматриваемом исследовательском проекте.

В большинстве случаев нейронные сети учатся с помощью алгоритма обратного распространения ошибок. Для использования этого алгоритма требуется наличие репрезентативной выборки данных, предварительно подготовленной для этого процесса, при условии, что для каждого элемента выборки известен результат.

Проблема заключается в том, что данных в исходной задаче недостаточно для нормального использования аппарата нейронных сетей, в связи с чем работать придется с той выборкой, которая имеется. Для искусственного увеличения выборки используется несколько методов. Первый метод увеличения количества данных, который был применен в проекте, — это кросс-валидация по блокам. Это происходит следующим образом: при стандартном обучении выборка разделяется на два блока. На одном нейронная сеть обучается, а на втором тестируется. При кросс-валидации она обучается на каждом из блоков и тестируется на противоположном, что позволяет удвоить количество данных для обучения. Второй метод — аугментация данных. Ее можно охарактеризовать как бесконтрольное искусственное изменение данных с целью получения псевдоподобных элементов. Эти методы применяются в нескольких блоках программного комплекса.

С учетом того, что сеть глубокая, а имеющаяся выборка небольшая по объему, возникла проблема переобучения сети. Ее суть заключается в том, что сеть запоминает исходную выборку целиком и не нацелена на получение средних значений. Данная проблема решалась с помощью добавления дополнительных промежуточных слоев «выбивание нейронов» и «нормализация слоев искусственной нейронной сети» в сети между уже имеющимися, а также с помощью регуляризации весовых коэффициентов.

Помимо этого, для смягчения последствий двух вышеуказанных проблем был использован автокодировщик. Это однослойная искусственная нейронная сеть, предназначенная для уменьшения шума данных, которая, как побочный эффект, всегда выдает разные результаты даже на одном и том же входном значении, что уменьшает вероятность переобучения и помогает слегка увеличить размер выборки.

ЗАКЛЮЧЕНИЕ

В данном диссертационном проекте рассматривается процесс автоматизации распознавания раковых клеток на микропрепарате при гистологическом исследовании. В рамках исследования проекта был разработан программно-аппаратный модуль для анализа и автоматической классификации клеток Рид – Березовского – Штернберга. Разработанный программный модуль использует несколько шагов для распознавания раковых клеток: предварительная подготовка, сегментация и классификация заболеваний. Также рассматриваются различные подходы для классификации и сегментации раковых клеток, а также производится оценка эффективности их использования.

В целом, получены удовлетворительные результаты сегментации и классификации раковых клеток. Результаты работы реализованных в программном модуле алгоритмов в большинстве случаев не уступают по функциональности уже существующим программным обеспечениям. Также предлагается способ улучшения качества обучаемой сети на малом объеме данных, используя принцип предварительной подготовки и аугментации экспериментальных данных. Данный способ удовлетворительно зарекомендовал себя в проведенных тестах. Помимо этого, в известных алгоритмах произведены улучшения, направленные на повышение скорости и качества их работы. Для повышения производительности применена нормализация слоев нейронной сети с помощью метода уменьшения измерений тренировочных данных.

В результате разработки цель проекта достигнута. Создан и улучшен программно-аппаратный модуль. На основании произведенного исследования написаны исследовательские работы [1–А] [2–А], опубликованные на международных научных конференциях.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1–A. Uladzislau Punko, Natallia Volorova, Uladzimir Prihodko. Morphological diagnosis of Reed – Sternberg cells in case of Hodgkin's lymphoma with artificial neural networks // МЕДЭЛЕКТРОНИКА-2018—2018—Режим доступа: https://www.bsuir.by/m/12_100229_1_132511.pdf

2–A. Uladzislau Punko, Natallia Volorova, Uladzimir Prihodko. Recognition of Reed – Sternberg cells in case of Hodgkin's lymphoma // Pattern Recognition and Infoformation Processing — 2019 — Режим доступа: https://doi.org/10.1134/S1054661820010125

3–A.Darya Filatova, Charles El-Nouty, Uladzislau Punko. HIGH-THROUGHPUT DEEP LEARNING ALGORITHM FOR DIAGNOSIS AND DEFECTS CLASSIFICATION OF WATERPROOFING MEMBRANES // International Journal for Computational Civil and Structural Engineering — 2020 — Режим доступа: https://doi.org/10.22337/2587-9618-2020-16-2-26-38