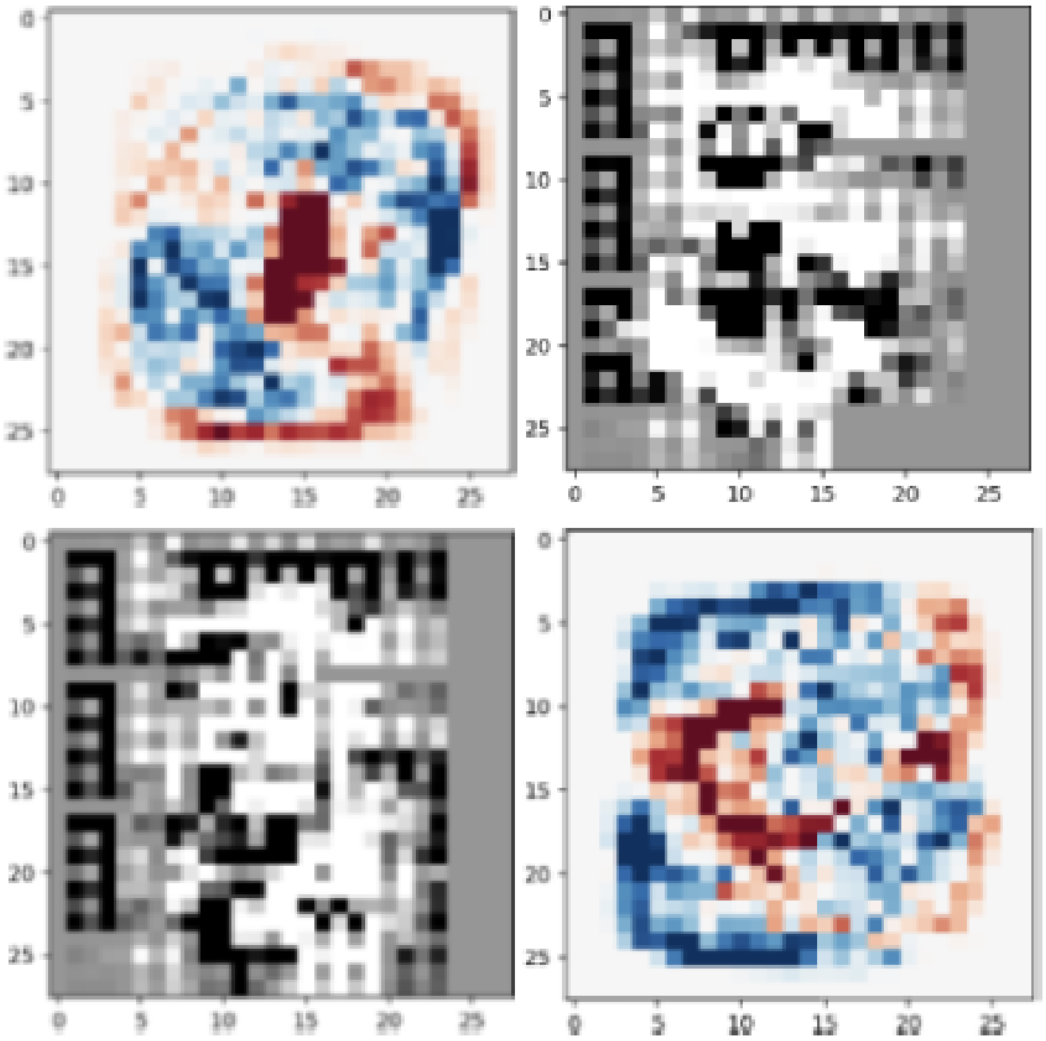
ИТМО

И. Ю. Попов, А. Я. Бучаев, Д. А. Есипов

ВАЛИДАЦИЯ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА



Санкт-Петербург
2024

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

УНИВЕРСИТЕТ ИТМО

И. Ю. Попов, А. Я. Бучаев, Д. А. Есипов

ВАЛИДАЦИЯ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

ЛАБОРАТОРНЫЙ ПРАКТИКУМ

РЕКОМЕНДОВАНО К ИСПОЛЬЗОВАНИЮ В УНИВЕРСИТЕТЕ ИТМО
по направлениям подготовки
10.04.01, 23.04.03, 11.04.03
в качестве учебно-методического пособия для реализации основных
профессиональных образовательных программ высшего образования
магистратуры

ИТМО

Санкт-Петербург
2024

Попов И. Ю., Бучаев А. Я., Есипов Д. А., Валидация систем искусственного интеллекта – СПб: Университет ИТМО, 2024. – 29 с.

Рецензент(ы):
Комаров И.И., к.ф.-м.н., доцент ФБИТ, Университет ИТМО.

Изложены основы современных способов и методов валидации и верификации систем искусственного интеллекта (ИИ), таких как валидация обучающей выборки, анализ интерпретации моделей ИИ, верификация моделей нейронных сетей. Особое внимание уделено интерпретации моделей нейронных сетей, что позволит изучить поведение модели и ее принципы принятия решений. Лабораторный практикум предназначен для студентов, обучающихся по программе магистерской подготовки по направлениям 10.04.01 – Информационная безопасность, 23.04.03 – Эксплуатация транспортно-технологических машин и комплексов, 11.04.03 – Конструирование и технология электронных средств, дисциплина «Валидация систем искусственного интеллекта».

ИТМО

Университет ИТМО – ведущий вуз России в области информационных и фотонных технологий, один из немногих российских вузов, получивших в 2009 году статус национального исследовательского университета. С 2013 года Университет ИТМО – участник программы повышения конкурентоспособности российских университетов среди ведущих мировых научно-образовательных центров, известной как проект «5 в 100». Цель Университета ИТМО – становление исследовательского университета мирового уровня, предпринимательского по типу, ориентированного на интернационализацию всех направлений деятельности.

Оглавление

Введение

Системы искусственного интеллекта (ИИ) имеют широкое применение: их повсеместно используют в экономике, финансах, промышленности, логистике, медицине, кинематографе и даже изобразительном искусстве, психологии и т.д. ИИ используются в системах автоматизированного управления, системах принятия решений, системах, связанных с обработкой естественного языка, распознаванием образов, речи, анализом данных, логистикой и других практически значимых задачах обработки данных.

Однако ошибки при решении таких задач могут привести к тяжким последствиям, связанным с серьезными экономическим и экологическими ущербами и даже угрозой для жизни и здоровья людей. Поэтому доверие к системам ИИ является важнейшим условием, определяющим возможность их применения.

В результате освоения дисциплины магистрант приобретает знания методов планирования вычислительного эксперимента, формирование обучающих и тестовых выборок. Умения использовать инструменты и библиотеки для валидации данных. Навыки работы по созданию валидационных данных для систем искусственного интеллекта.

В лабораторном практикуме объединены лабораторные работы по анализу, валидации и визуализации обучающей выборки, оценке модели нейронной сети и составлению ее паспорта. Большинство лабораторных работ рассчитаны на создание обучающимися некоторого программного обеспечения на языке программирования Python.

Лабораторный практикум предназначен для студентов, обучающихся по программе магистерской подготовки по направлениям 10.04.01 – Информационная безопасность, 23.04.03 – Эксплуатация транспортно-технологических машин и комплексов, 11.04.03 – Конструирование и технология электронных средств, дисциплина «Валидация систем искусственного интеллекта».

Все задания могут быть выполнены в бесплатном облачном сервисе Google Colaboratory или любой другой среде разработки, поддерживающей Python, выбор среды разработки не влияет на ход выполнения лабораторных работ.

Постановка каждой задачи лабораторного практикума предельно коротка, однако для понимания задачи необходимо знать материал соответствующего раздела курса «Валидация систем искусственного интеллекта», а для написания вспомогательных скриптов – уметь программировать на Python.

При необходимости более глубокого изучения какой-либо темы обучающиеся могут обратиться к литературе, представленной в практикуме после каждой лабораторной работы.

Данный лабораторный практикум предназначен для закрепления знаний, приобретения умений, практических навыков и формирования компетенций обучающихся по валидации систем искусственного интеллекта. Практикум включает 4 лабораторные работы, предусматривающих выполнение заданий в период СРС. Каждая лабораторная работа определяется целью, задачами, теоретической частью, порядком выполнения работы. Требования к отчету и защите отчета по лабораторной работе изложены в конце данного практикума.

Пререквизиты и результаты обучения

Для успешного выполнения лабораторного практикума и освоения материала дисциплины студенты должны иметь подготовку по следующим дисциплинам:

- Основы программирования Python.
- Анализ данных и основы машинного обучения.

По итогам освоения дисциплины студент сможет:

- проводить валидацию обучающего набора данных;
- верифицировать и валидировать модели нейронных сетей обнаружения изображений;
- проводить анализ программного кода формирования модели нейронной сети на предмет аномалий и корректировки гиперпараметров;
- использовать Python для написания вспомогательных инструментов интерпретации и визуализации слоев нейронной сети;
- использовать существующие средства интерпретации слоев нейронной сети;
- использовать Python для написания вспомогательных инструментов поиска и проверки гипотез, основанных на анализе графических интерпретаций.

Программа лабораторного практикума

Практикум включает четыре логически и информационно связанные лабораторные работы, последовательно реализующие основные этапы процесса валидации систем искусственного интеллекта (Рисунок 1).

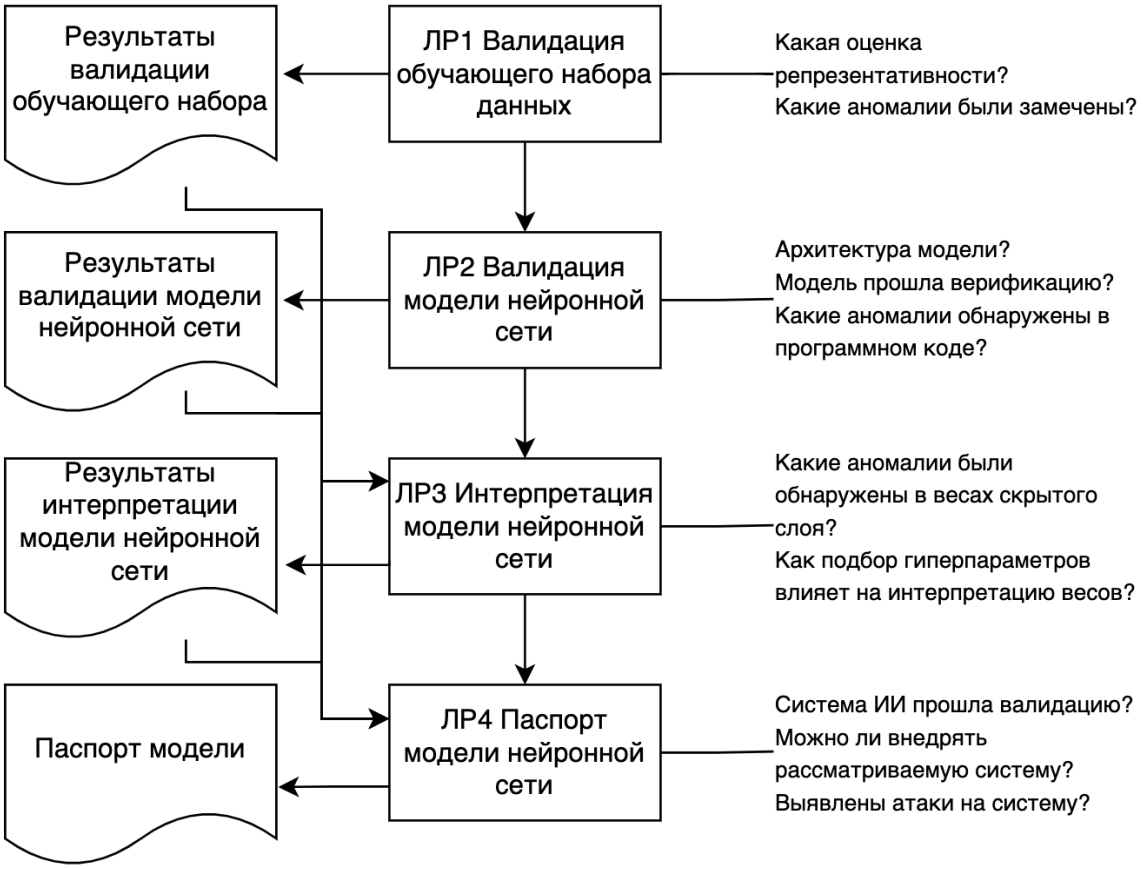


Рисунок 1. Концепция последовательной реализации лабораторных работ практикума

Лабораторная работа 1. Валидация обучающего набора данных

Цели работы – изучение основных методов валидации обучающего набора данных; получение навыков анализа и валидации обучающей выборки; получение навыков визуализации полученных данных в различных графических форматах.

Помимо составления краткой сводки полученной информации и оформления отчета, лабораторная работа состоит из четырех частей: в первой части описывается представленный обучающий набор данных, во второй части производится поиск аномальных элементов, в третьей части оценивается репрезентативность данных, в четвертой части визуализируются полученные результаты.

Лабораторная работа 2. Валидация модели нейронной сети классификации изображений

Цели работы – изучение основных методов валидации целевой модели; получение навыков анализа и описания модели.

Работа состоит из четырех частей, помимо составления краткой сводки полученной информации и оформления отчета по лабораторной работе. В первой части описывается архитектура модели нейронной сети классификации изображений согласно варианту, во второй части производится верификация модели с помощью предоставленного тестового набора данных, в третьей части осуществляется проверка представленного программного кода, в четвертой части оценивается качество работы модели.

Лабораторная работа 3. Интерпретация модели нейронной сети классификации изображений

Цели работы – получение навыков исследования механизма принятия решений моделей классификации изображений и формирования представления об интерпретации обрабатываемых моделью данных.

Работа состоит из трех частей, помимо составления краткой сводки полученной информации и оформления отчета по лабораторной работе. В первой части исследуются инструменты интерпретации и визуализации элементов нейронной сети, во второй части производится интерпретация весов пикселей модели в виде изображения, в третьей части исследуется интерпретация весов пикселей на предмет аномалий.

Лабораторная работа 4. Паспорт модели нейронной сети классификации изображений. Заключающая характеристика

Цели работы – получение навыков сводного анализа полученных результатов; написание заключающей характеристики и составление паспорта модели.

Работа состоит из трех частей, которые объединяют результаты всех предыдущих работ, помимо оформления отчета по лабораторной работе. В первой части составляется паспорт модели. Во второй части разрабатывается заключающая характеристика. В третьей части проводится защита результатов.

Требования к оформлению отчета по лабораторным работам

Общие требования

- Отчет выполняется в виде самодостаточного документа. Материал, изложенный в отчете, должен быть понятен без дополнительных комментариев со стороны исполнителей.
- Отчет выполняется в виде текстового документа в соответствии с ГОСТ 7.32-2017 и представляется в электронном виде.
- В отчете должен(-ы) быть полностью представлен(-ы) вспомогательный(-ые) скрипт(-ы) на Python, которые были написаны для выполнения задач лабораторной работы.
- Листы отчета должны быть пронумерованы, кроме первого листа, который имеет номер 1 и считается титульным.
- Если отчет содержит большое количество листов, рекомендуется добавлять лист с содержанием отчета (разделы и номера листов).

При оформлении отчетных документов целесообразно ориентироваться на «Требования к выпускным квалификационным работам», утвержденным решением Ученого совета Университета ИТМО (на момент подготовки рекомендаций Версия 4 от «29» ноября 2022 г., код идентификации документа ЛНАОБУЧ-СМК-03-05-2022).

Применительно к оформлению отчета наиболее существенными являются следующие рекомендации этого документа:

- Шрифт - Times New Roman, не менее 12 пт. (рекомендован 14 пт.).
- Межстрочный интервал - 1,5.
- Каждая новая «красная» строка должна иметь абзацный отступ 1,25 см.
- Основной текст – выравнивание «по ширине».
- Рисунки и подрисуночные подписи – выравнивание «по центру» без абзацного отступа.
- Ширина полей: левого 3 см, правого 1 см, верхнего и нижнего по 2 см.
- Объем отчета не должен превышать 30 страниц при оформлении шрифтом Times New Roman 14 пт.

Примеры и правила оформления списка терминов, иллюстраций, формул, ссылок, списка литературы и приложений определены упомянутым документом и требованиями ГОСТ 7.32– 2017.

Содержание отчета

Отчет должен содержать следующие разделы:

1) Титульный лист.
2) Тема, цель и задачи работы.
3) Содержательная часть.
4) Выводы.

Примеры выполнения лабораторных работ №1 и №2 представлены в Приложениях 1 и 2.

Тема, цель и задачи работы

В данном разделе должны быть указаны тема и цель работы. Их необходимо скопировать из данного пособия. При необходимости цель и задачи уточняются преподавателем при выдаче задания. Также должны быть указаны задачи, решаемые в процессе выполнения лабораторной работы для достижения поставленной цели.

Содержательная часть

В данном разделе должны быть представлены и описаны шаги для достижения поставленной цели. Для этого содержательная часть условно делится на теоретическую и экспериментальную.

Теоретическая часть отчета по лабораторной работе должна включать в себя всю необходимую информацию о предметной области, к которой относятся основные закономерности, принципы работы программных компонентов, используемых при ее выполнении, и т.п.

Экспериментальная часть лабораторной работы является обязательной и не может быть опущена. В ней отражаются персональные результаты обучающегося, полученные им при выполнении лабораторной работы. Этими результатами являются таблицы и изображения, полученные в процессе решения сформулированных задач, а также характеристики и аналитические заключения.

Также должны быть представлены отдельно исходные данные программного кода, использованные обучающимся для проверки реализованных в ходе практической работы компонентов, в том числе в виде ссылки на Интернет-ресурс.

Если обучающийся не завершил цикл выполнения программы, должно быть проведено ее окончательное редактирование. В экспериментальную часть отчета по лабораторной работе необходимо включить пояснения и комментарии к написанному программному коду (если написание программного кода является одной из задач лабораторной работы).

Выводы

Выводы должны быть кратко изложены в виде списка и отражать наиболее важные аспекты лабораторной работы. В конце отчета студент должен привести список использованной при подготовке отчета и процитированной литературы (в том числе ссылки на стандарты, руководства пользователя и прочую техническую документацию). Текст отчета по лабораторной работе может являться планом ответа при защите лабораторной работы.

Правила оформления текста отчета по лабораторной работе

Отчет должен быть представлен в формате PDF. На рисунках в отчете могут быть представлены блок-схемы, графики, снимки экранов виртуального окружения, а также прочая графическая информация. Рисунки должны быть четкими и легко читаемыми.

Листинги программного кода должны оформляться как скриншоты окна редактора, либо непосредственно в тексте отчета. В обоих случаях должны быть соблюдены следующие правила:

1) должны использоваться моноширинные шрифты;
2) должны быть пронумерованы строки программы (нумерация должна начинаться с единицы);
3) должна быть включена «подсветка» синтаксических конструкций используемого языка программирования.

Исходные коды должны быть снабжены комментариями. На рисунки, листинги и таблицы обязательно должны присутствовать ссылки в тексте. Листинги, таблицы и рисунки должны быть снабжены подписями, отражающими их содержание.

Защита лабораторной работы

Во время защиты лабораторной работы обучающемуся следует ясно и четко изложить ключевые этапы решения поставленных задач, а также ответить на дополнительные вопросы, заданные преподавателем в процессе защиты отчета. Критерии, по которым оценивается защита лабораторной работы, представлены в таблице 1.

Таблица 1 – Критерии оценивания защиты лабораторной работы

№ п/п	**Критерий**	**Оценка (уровень)**		
		Высокий	**Средний**	**Низкий**
1	Корректность оформления и оригинальность отчета	2	1,5	1
2	Качество устного представления результатов работы	2	1,5	1
3	Понимание и воспроизведение представленного программного кода	2	1,5	1
4	Правильность и полнота аналитического заключения достигнутого результата	2	1,5	1
5	Качество ответов на дополнительные вопросы	2	1,5	1
	Итого баллов:	**10**	**7,5**	**5**

Рекомендуется ознакомиться со списком общей литературы в целях получения дополнительных теоретических и практических знаний, а также развития каузального мышления при анализе систем искусственного интеллекта.

Список общей рекомендуемой литературы:

1) Ватьян А.С., Гусарова Н.Ф., Добренко Н.В. Системы искусственного интеллекта. – СПб: Университет ИТМО, 2022. – 186 с.
2) Леонов Ф.В., Челпанов А.Д., Югансон А.Н., Программирование на языке Python для решения задач информационной безопасности. Методические указания по выполнению практических работ.– СПб: Университет ИТМО, 2021. – 36 с.
3) Орлов Г. М., Игнатьева О. А., Васин А. Г., Низомутдинов Б. А. Современные методы обработки и анализа данных. – СПб.: Университет ИТМО, 2021. – 147 с.

Лабораторная работа № 1

Валидация обучающего набора данных

Цели работы:

Изучение основных методов валидации обучающего набора данных; получение навыков анализа и валидации обучающей выборки; получение навыков визуализации полученных данных в различных графических форматах.

Задачи:

1) описать представленный набор данных;
2) произвести поиск аномальных элементов набора данных;
3) оценить репрезентативность данных;
4) визуализировать полученные результаты;
5) составить краткую сводку полученной информации.

Ход работы:

Для выполнения данной работы необходимо скачать набор данных согласно варианту (https://disk.yandex.ru/d/YJiSKAt5c8AQvg) и ознакомиться с прилагаемым набором данных.

Описательной характеристикой обучающего набора данных является сжатая характеристика, представленная в виде графиков, таблиц, схем и числовых значений, характеризующих обучающий набор данных, например, количество представителей того или иного класса.

Необходимо написать вспомогательные скрипты на языке программирования Python, которые решают перечисленные задачи.

Под аномальностью элементов обучающего набора можно понимать, например, изображения, имеющие несоответствующие разрешение, цветовое пространство, размерность и тд. Изображения также могут содержать специфичные варианты написания рукописных цифр, которые в некоторых случаях негативно влияют на обучение модели, повышая отклонения в весах скрытых слоев полносвязной сети. Данное ограничение обусловлено спецификой используемого типа нейронной сети. При использовании иных типов нейронных сетей необходимо учитывать соответствующие особенности обработки входных изображений.

Одним из способов выделения элементов, содержащих артефакты является снижение размерности и поиск аномалий в двумерном или трехмерном пространствах, например, использование алгоритма снижения размерности UMAP (рисунок 1).

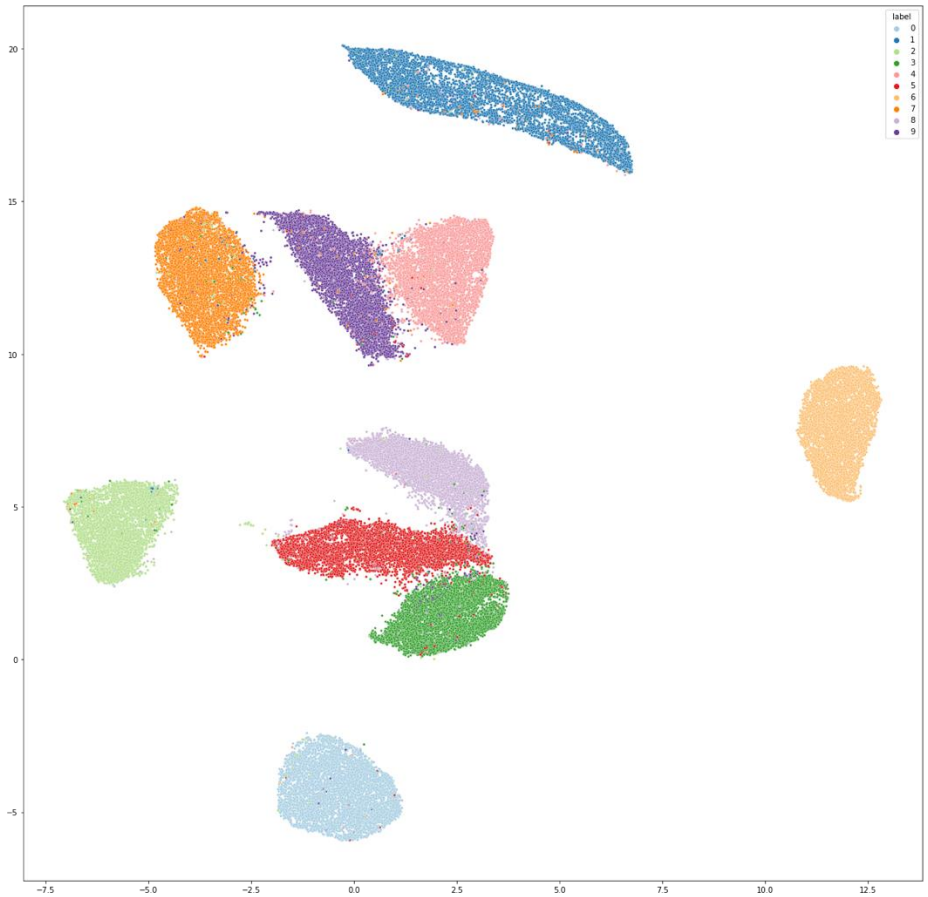


Рисунок 1 – Распределение классов при визуализации результата работы алгоритма UMAP

Репрезентативность модели отражает полноту представления данных и влияет на качество классификации модели. Необходимо произвести качественную и количественную оценку репрезентативности.

Под качественной репрезентативностью понимается распределение представителей классов, стремящееся к равномерному, то есть каждый элемент исследуемого набора данных имеет одинаковую вероятность выбора при случайном отборе.

Количественная репрезентативность определяет необходимое количество представителей классов, указывает, является ли число элементов выборки достаточным для представления генеральной совокупности с учетом некоторой ошибки, и вычисляется по формуле 1.

$$n = \frac{t^2 * p * q}{\Delta^2}, \tag{1}$$

где t – доверительный коэффициент, показывающий, какова вероятность того, что размеры показателя не будут выходить за границы предельной ошибки; p – доля единиц наблюдения, обладающих изучаемым признаком; $q = 1 - p$ — доля единиц наблюдения, не обладающих изучаемым признаков; Δ — допустимая ошибка выборки.

Для визуализации полученных данных рекомендуется использование программных инструментов, например, библиотеки matplotlib. Столбчатые и круговые диаграммы обеспечат хорошее восприятие количества представителей между классами.

Список рекомендуемой литературы:

1) V. Varkarakis and P. Corcoran, "Dataset Cleaning — A Cross Validation Methodology for Large Facial Datasets using Face Recognition," 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX), Athlone, Ireland, 2020, pp. 1-6, doi: 10.1109/QoMEX48832.2020.9123123.
2) Пылов, П. А. Фундаментальные типы кросс-валидации для оценки качества моделей машинного и глубокого обучения / П. А. Пылов, О. А. Ивина // Россия молодая : Сборник материалов XIII Всероссийской научно-практической конференции с международным участием, Кемерово, 20–23 апреля 2021 года / Редколлегия: К.С. Костиков (отв. ред.) [и др.]. – Кемерово: Кузбасский государственный технический университет имени Т.Ф. Горбачева, 2021. – С. 31520.1-31520.7. – EDN NYULSQ.
3) AltexSoft, Data Quality Management: Roles, Processes, Tools [Электронный ресурс] .- https://www.altexsoft.com/blog/data-quality-management-and-tools/ (дата обращения: 23.01.2023)
4) Хабр, О важности датасета и о том, как сделать его лучше. Опыт нашей компании [Электронный ресурс] .- https://habr.com/ru/post/678808/ (дата обращения: 23.01.2023)
5) GeekBrains, Датасет: виды, применение, набор лучших [Электронный ресурс] .- https://gb.ru/blog/dataset/ (дата обращения: 23.01.2023)
6) Loginom, Репрезентативность выборочных данных [Электронный ресурс] .- https://loginom.ru/blog/representativity (дата обращения: 23.01.2023)

Лабораторная работа № 2

Валидация модели нейронной сети классификации изображений

Цели работы

Изучение основных методов валидации целевой модели; получение навыков анализа и описания модели.

Задачи:

1) описать архитектуру модели нейронной сети классификации изображений согласно варианту;
2) произвести верификацию модели с помощью предоставленного набора данных;
3) осуществить проверку представленного программного кода;
4) оценить качество работы модели;
5) составить краткую сводку полученной информации.

Ход работы:

Для выполнения данной работы необходимо скачать необходимые файлы согласно варианту (https://disk.yandex.ru/d/YJiSKAt5c8AQvg), выполнить описание модели согласно варианту.

Для получения детальной характеристики модели можно анализировать прилагаемый исходный код или использовать вспомогательные методы библиотек, например, tensorflow позволяет получить структуру модели (рисунок 2) и графическое представление в виде блоков (рисунок 3).

```
Model: "sequential_8"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 flatten_8 (Flatten)         (None, 784)               0

 dense_16 (Dense)            (None, 16)                12560

 dropout_8 (Dropout)         (None, 16)                0

 dense_17 (Dense)            (None, 10)                170

=================================================================
Total params: 12730 (49.73 KB)
Trainable params: 12730 (49.73 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Рисунок 2 – Описание слоев модели машинного обучения

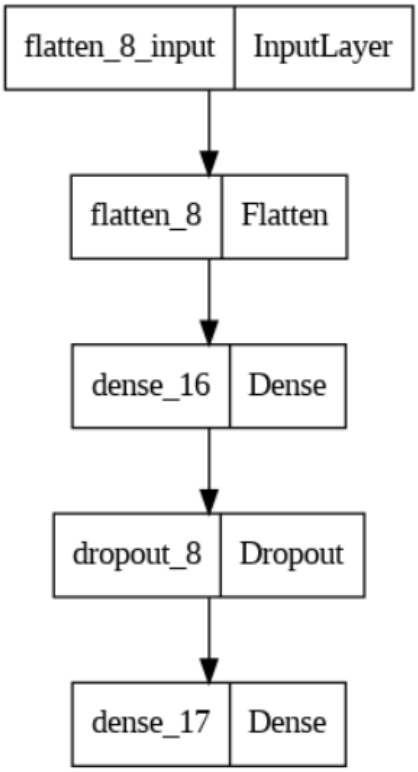


Рисунок 3 – Архитектура модели машинного обучения

Иным инструментом визуализации является кроссплатформенная утилита Netron от автора Lutz Roeder, которая позволяет визуализировать ноды моделей машинного обучения и их взаимосвязи (рисунок 4). Платформа доступа для скачивания, а также имеет веб-ресурс (https://netron.app/), который предоставляет тот же функционал. Для установки необходимо загрузить файлы с помощью команды *pip install netron*.

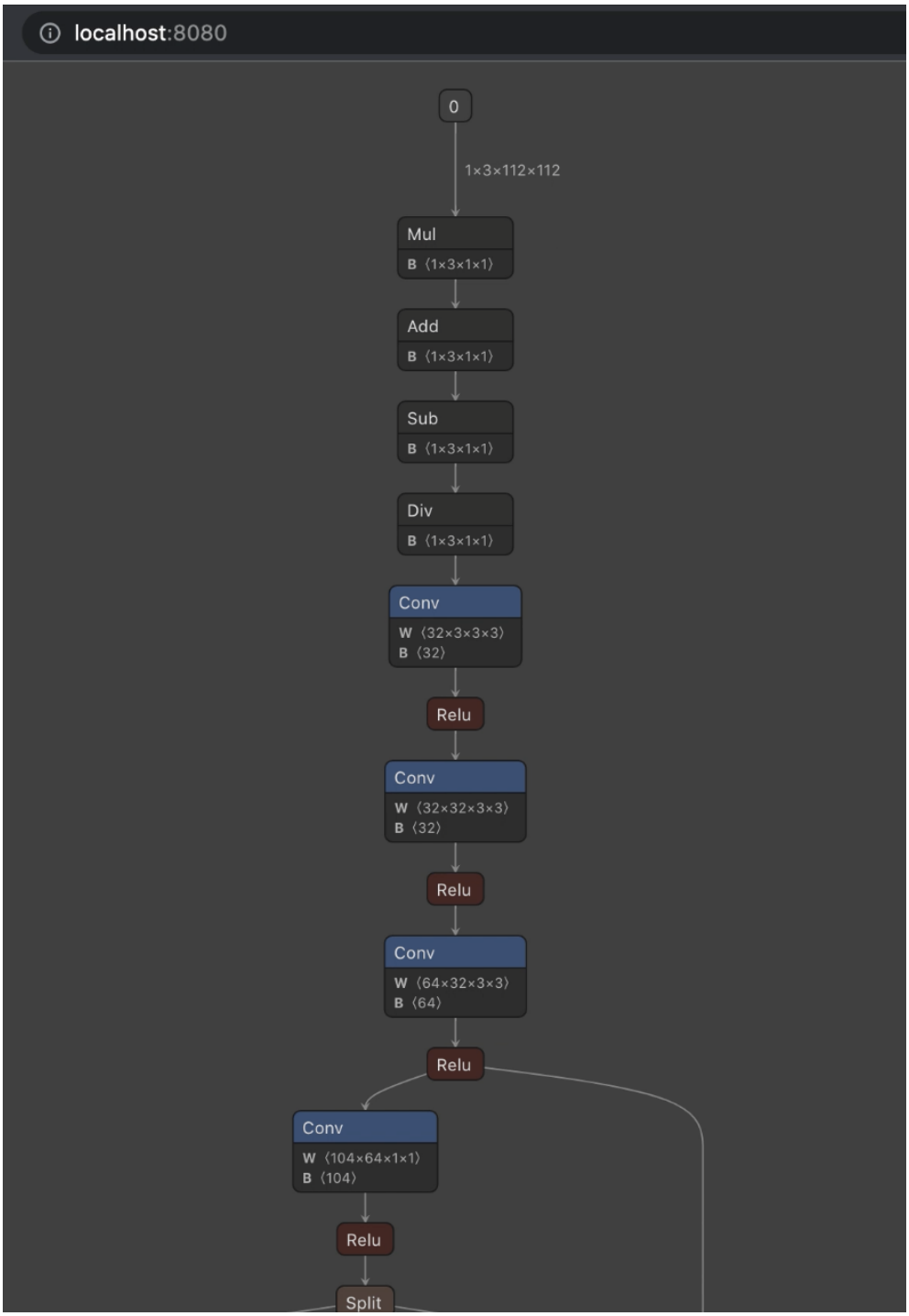


Рисунок 4 – Интерактивная визуализация взаимосвязей нод модели машинного обучения

Представленный инструмент позволяет в интерактивном режиме просматривать размерности и типы входных и выходных данных, указывает конфигурируемые параметры используемых в модели машинного обучения нод, в том числе веса скрытых слоев и иные матрицы преобразования.

Проведение верификации подразумевает использование прилагаемого набора данных для получения оценок качества работы модели, например, точность (2) и полнота (3).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (2)$$

$$Recall = \frac{TP}{TP + FN}, \quad (3)$$

Стоит уточнить, что существуют иные методы и оценки моделей машинного обучения, например, для моделей распознавания лиц используются показатели FRR (False Rejection Rate) — доля неправильно отклонённых легитимных-попыток; FAR (False Acceptance Rate) — доля неправильно принятых нелегитимных-попыток.

Проверка программного кода предполагает анализ построения модели, оценку подхода формирования выборок, исследования установленных гиперпараметров модели и их влияния при иных значениях. К исследуемым параметрам можно отнести, например, batch_size, learning rate и тд. Также к исследуемым параметрам следует отнести конфигурацию слоев модели, например, значение слоя dropout или используемые функции активации.

Список рекомендуемой литературы:

1) Пылов, П. А. Фундаментальные типы кросс-валидации для оценки качества моделей машинного и глубокого обучения / П. А. Пылов, О. А. Ивина // Россия молодая : Сборник материалов XIII Всероссийской научно-практической конференции с международным участием, Кемерово, 20–23 апреля 2021 года / Редколлегия: К.С. Костиков (отв. ред.) [и др.]. – Кемерово: Кузбасский государственный технический университет имени Т.Ф. Горбачева, 2021. – С. 31520.1-31520.7. – EDN NYULSQ.
2) OSP, Валидация автономных систем [Электронный ресурс] .- https://www.osp.ru/os/2019/04/13055222 (дата обращения: 23.01.2023)
3) Хабр, Валидация моделей машинного обучения [Электронный ресурс] .- https://habr.com/ru/company/glowbyte/blog/569970/ (дата обращения: 23.01.2023)

Лабораторная работа № 3

Интерпретация модели нейронной сети классификации изображений

Цели работы:

Получение навыков исследования механизма принятия решений моделей классификации изображений; формирование представления об интерпретации обрабатываемых моделью данных.

Задачи:

1) исследовать инструменты интерпретации и визуализации элементов нейронной сети;
2) произвести интерпретацию весов пикселей модели в виде изображения;
3) исследовать интерпретацию весов пикселей на предмет аномалий;
4) составить краткую сводку полученной информации.

Ход работы:

Для выполнения лабораторной работы возможно использование следующих инструментов:

1) tensorboard;
2) sklearn.

В ходе выполнения работы требуется скачать необходимые файлы согласно варианту (https://disk.yandex.ru/d/YJiSKAt5c8AQvg). Анализ скрытых слоев моделей обработки изображений позволяет выявить смещения фокуса модели от предполагаемого, визуальные артефакты, которые понижают точность модели, либо могут быть использованы для реализации атак модификации модели (https://bdu.fstec.ru/threat/ubi.221). Для проведения ручного анализа возможно исследование интерпретаций весов скрытых слоев модели. Интерпретации бывают разных размерностей и форматов, наиболее оптимальными в рамках исследования моделей обработки изображений являются графические интерпретации, которые представляют двумерные или трехмерные объекты.

Одним из способов получения графических интерпретаций весов скрытых слоев является метод закрашивания окна. Суть метода заключается в модификации анализируемого изображения путем закрашивания небольшой области, затем модифицированное изображение анализируется нейронной сетью и вычисляется степень принадлежности к определенному классу (рисунок 5). Данный процесс является итеративным и должен покрывать все области изображения для получения информативной интерпретации. При закрашивании наиболее чувствительных участков

изображения степень принадлежности падает, что говорит о мере влияния данного участка на принятие решения классификации.

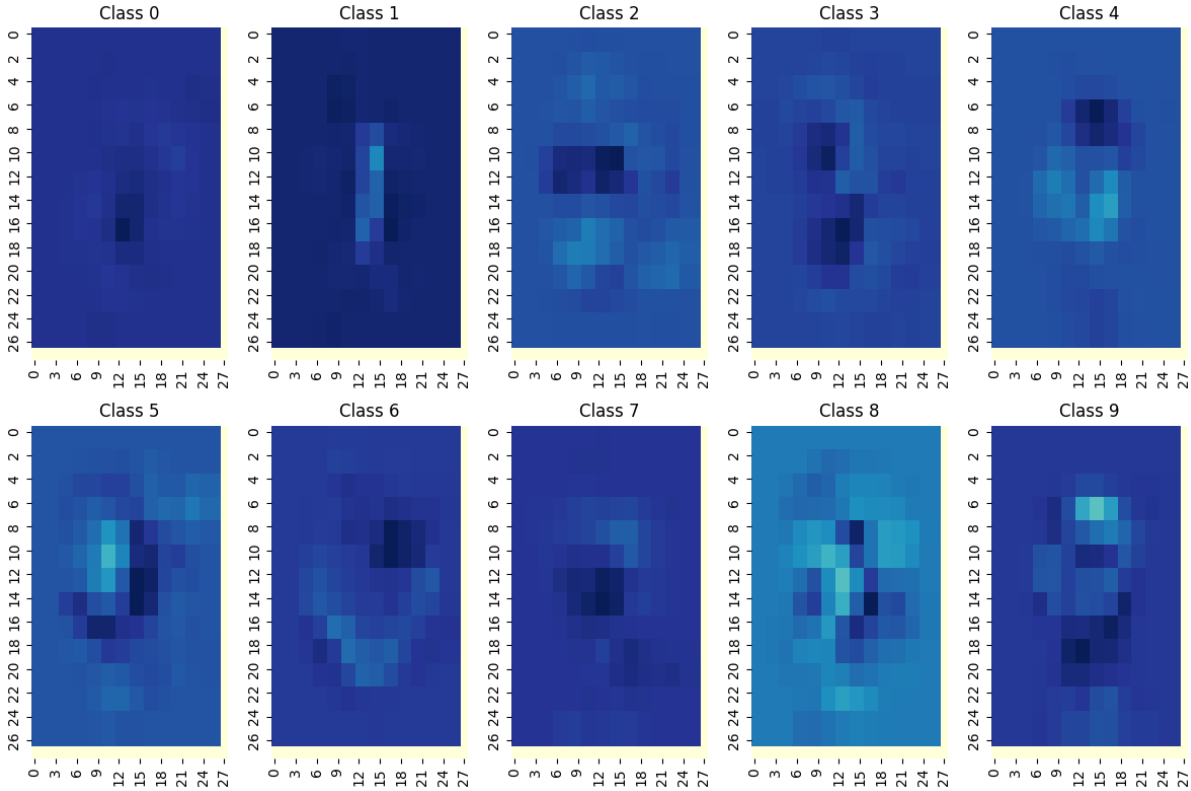


Рисунок 5 – Выявление чувствительных участков изображений методом закрашивания

Также существуют инструменты автоматизированной визуализации весов скрытых слоев, например, инструмент tensorboard (https://www.tensorflow.org/tensorboard/image_summaries?hl=ru), который предоставляет краткую информацию о модели на основе использованных логов и файлов модели. С его помощью можно, например, получить распределение смещения (рисунок 6) для анализа обучающей выборки и тестовой выборки на предмет статистических различий. Иным инструментом построения моделей машинного обучения, содержащим функционал автоматизированной визуализации, является библиотека sklearn.

При исследовании графических интерпретаций разумно использовать статистические методы и инструменты. При обучении полносвязных моделей нейронных сетей возникают строгие принципы принятия решения модели, которые можно выявить при статистическом анализе всех элементов того или иного класса. Обуславливаются такие принципы строгим соответствием анализируемых пикселей и их весов.

Вспомогательные скрипты, выполняющие перечисленные задачи, необходимо писать на языке Python.

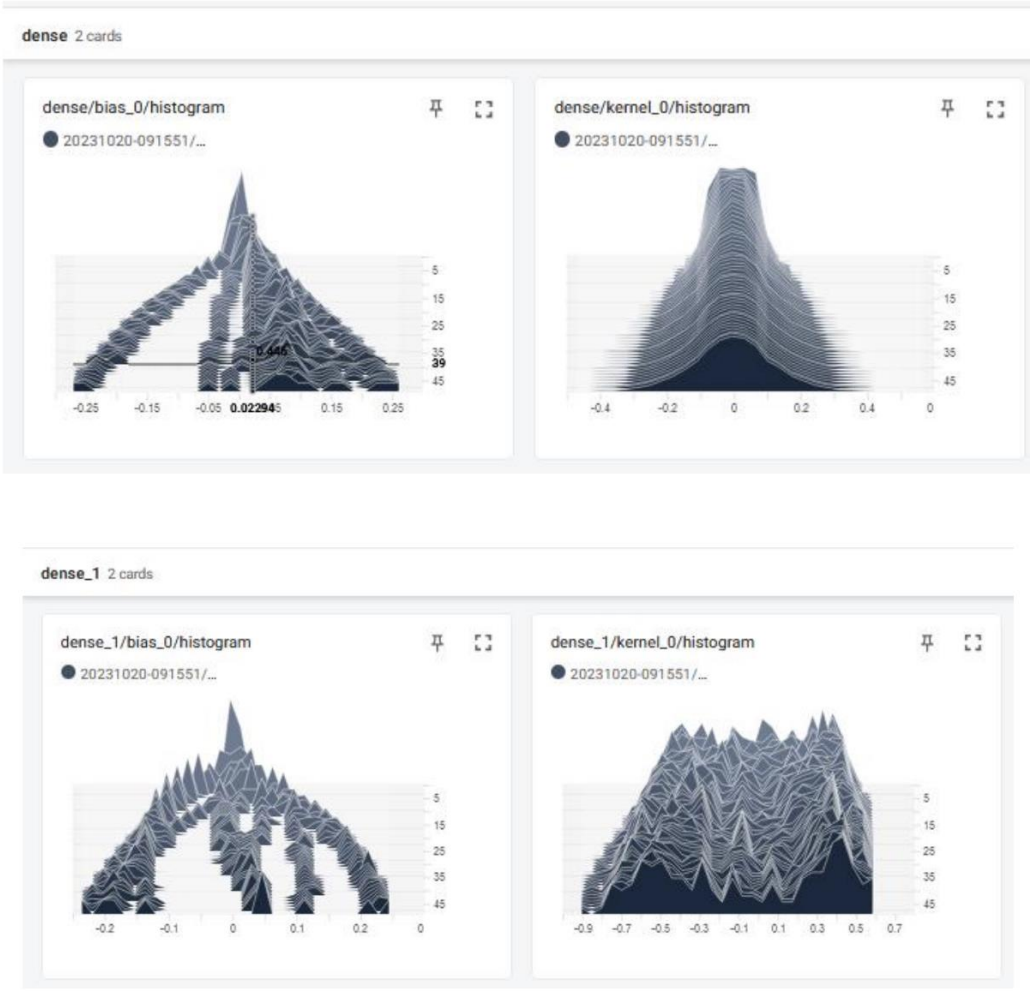


Рисунок 6 – Гистограммы распределения весов и смещений в двух скрытых слоях

Список рекомендуемой литературы:

1) Zeiler, M.D., Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8689. Springer, Cham. https://doi.org/10.1007/978-3-319-10590-1_53
2) Owen Shen, "Interpretability in ML: A Broad Overview", The Gradient, 2020.
3) Хабр, «Сделать красиво». Визуализация обучения с Tensorboard от Google [Электронный ресурс] .- https://habr.com/ru/post/349338/ (дата обращения: 23.01.2023)
4) NewTechAudit, Визуализация архитектуры и отдельных блоков нейросети с помощью Netron [Электронный ресурс] .- https://newtechaudit.ru/netron/ (дата обращения: 23.01.2023)

5) Хабр, Визуализация процесса обучения нейронной сети средствами TensorFlowKit [Электронный ресурс] .- https://habr.com/ru/post/342934/ (дата обращения: 23.01.2023)
6) StevenRush, Визуализация того, чему обучается сверточная нейронная сеть [Электронный ресурс] .- https://stevenrush.github.io/understanding-cnn/ (дата обращения: 23.01.2023)

Лабораторная работа № 4

Паспорт модели нейронной сети классификации изображений. Заключающая характеристика

Цели работы:

Получение навыков сводного анализа полученных результатов; написание характеристики модели и составление паспорта.

Задачи:

1) составить паспорт модели;
2) написать заключающую характеристику;
3) провести защиту результатов.

Ход работы:

Для выполнения лабораторной работы необходимо подготовить паспорт модели и итоговое заключение, содержащее взаимосвязанные результаты прошлых работ.

Данные документы имеют свободный формат, требуется информативное заключение, краткая сводка по всем полученным результатам. Подразумевается использование материалов предыдущих работ для формирования паспорта и заключения.

В заключении необходимо представить результат валидации модели, а также связать итоговую оценку с ранее установленными характеристиками. Результаты работ должны дополнять друг друга, так, например, возможно установить влияние гиперпараметров модели на качество интерпретации весов скрытых слоев.

В случае выявления ограничений или уязвимостей модели необходимо аргументировать на основе полученных результатов эти факты и оценить степень их влияния на работу модели. Процесс защиты представляет собой оценку полученных результатов и написанного программного кода.

Приложение 1

Пример выполнения Лабораторной работы №1

Отчет по лабораторной работе №1

По дисциплине: «Валидация систем искусственного интеллекта»

На тему: «Валидация обучающего набора данных»

Выполнил(-а):

Магистрант гр. N42101c Иванов И. И.

_____

Проверил(-а):

Доцент ФБИТ, к.т.н. Попов И. Ю.

_____

Санкт-Петербург
2024

Тема работы: Валидация обучающего набора данных.

Цели работы: изучение основных методов валидации обучающего набора данных; получение навыков анализа и валидации обучающей выборки; получение навыков визуализации полученных данных в различных графических форматах.

Задачи:

1) описать представленный набор данных;

2) произвести поиск аномальных элементов набора данных;

3) оценить репрезентативность данных;

4) визуализировать полученные результаты;

5) составить краткую сводку полученной информации.

Описание набора данных

Представленный набор данных является набором изображений. Данный набор содержит в себе совокупности пяти различных классов фруктов: яблоки, бананы, груши, манго и апельсины.

Суммарно в обучающем наборе данных 5000 изображений, каждый класс представляется 1000 экземплярами соответствующего фрукта.

Каждый экземпляр представляет собой изображение размером 30x30 пикселей в цветовом пространстве RGB.

Поиск аномальных элементов обучающего набора данных

Для обработки обучающего набора данных был написан программный код на языке Python (рисунок 1), который реализует поиск аномальных данных, а именно удаление аномальных экземпляров:

1) дубликаты;

2) изображения нестандартного размера;

3) изображения иного формата;

4) изображения в серых тонах.

Такие характеристики были выбраны, так как они влияют на репрезентативность обучающей выборки и корректность работы модели, а также считается, что цветовая компонента важна для классификации фруктов, следовательно, изображения в оттенках серого можно считать выбросами в выборке.

```python
# Импорт необходимых фреймворков для загрузки и проверки входных данных

from PIL import Image, ImageDraw
from utils import check_size, check_color, check_format, load_dataset


# Функция проверки валидности изображения по размеру, формату и цве
def validate_image(image):
    size = check_size(image)
    image_format = check_format(image, size)
    color = check_color(image)

    return size and image_format and color


# Проверка и удаление аномальных изображений, внутри датасета
def check_dataset(dataset):
    for i in range(len(dataset)):
        if not(validate_image(dataset[i])):
            del dataset[i]


def main():
    # Загужаем датасет наиболее потребляемых фруктов
    data = load_dataset('fruits')

    # Проверяем и удаляем аномальные изображения
    check_dataset(data)


if __name__ == '__main__':
    main()
```

Рисунок 1 – Поиск аномальных данных в обучающем наборе

Оценка репрезентативности данных

В рамках выполнения лабораторной работы необходимо произвести качественную и количественную оценку репрезентативности данных.

При качественной оценки важным фактором является наличие в выборке всех представленных классов с равной вероятностью выбора того или иного класса. С помощью инструментов визуализации на языке Python была получена круговая диаграмма, показывающая относительное содержание экземпляров исследуемых классов (рисунок 2).

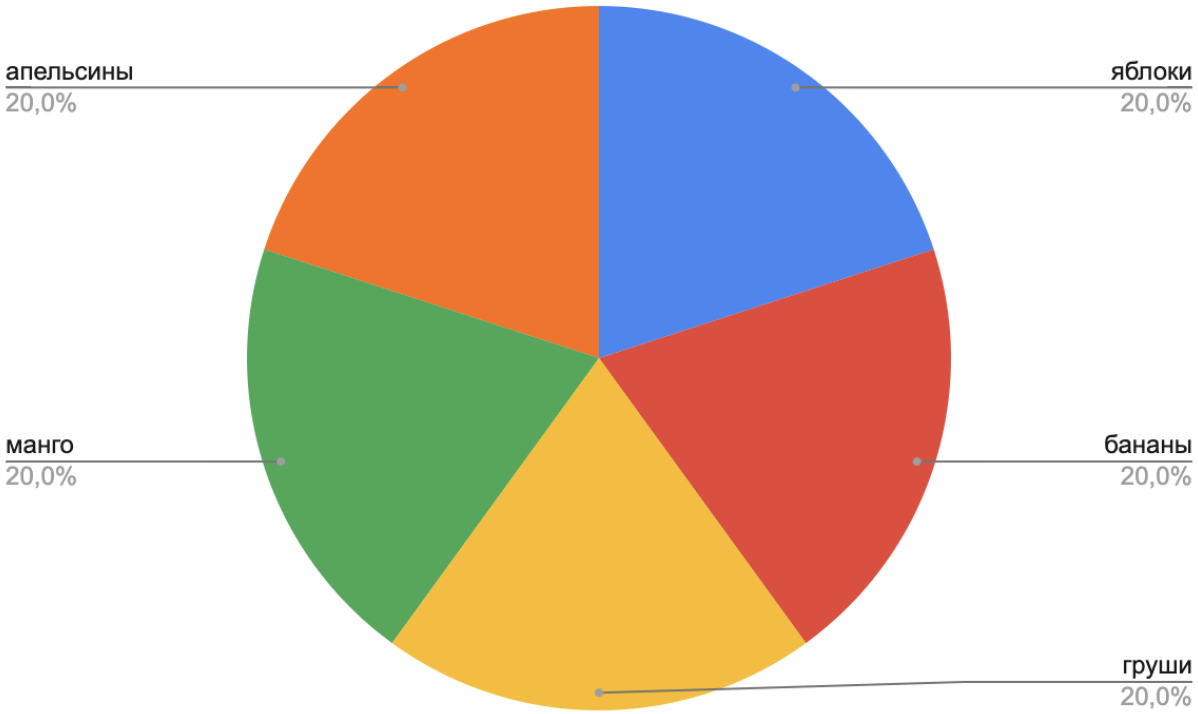


Рисунок 2 – Оценка качественной репрезентативности классов выборки

Количественная оценка вычисляется с помощью формулы 1:

$$n = \frac{t^2 * p * q}{\Delta^2}, \quad (1)$$

где t – доверительный коэффициент, показывающий, какова вероятность того, что размеры показателя не будут выходить за границы предельной ошибки; p – доля единиц наблюдения, обладающих изучаемым признаком; $q = 1 - p$ — доля единиц наблюдения, не обладающих изучаемым признаков; Δ — допустимая ошибка выборки.

Выводы

В ходе выполнения работы была получена оценка репрезентативности обучающего набора, согласно которой обучающий набор является репрезентативным, аномальные элементы в соответствии с предполагаемыми задачами и целями при использовании данного обучающего набора данных не были найдены.

Приложение 2

Пример выполнения Лабораторной работы №2

ИТМО

Отчет по лабораторной работе №2

По дисциплине: «Валидация систем искусственного интеллекта»

На тему: «Валидация модели нейронной сети классификации изображений»

Выполнил(-а):

Магистрант гр. N42101c Иванов И. И.

_____

Проверил(-а):

Доцент ФБИТ, к.т.н. Попов И. Ю.

_____

Санкт-Петербург
2024

Тема работы: валидация модели нейронной сети классификации изображений.

Цели работы: изучение основных методов валидации целевой модели; получение навыков анализа и описания модели.

Задачи:

1) описать архитектуру модели нейронной сети классификации изображений согласно варианту;
2) произвести верификацию модели с помощью предоставленного тестового набора данных;
3) осуществить проверку представленного программного кода;
4) оценить качество работы модели;
5) составить краткую сводку полученной информации.

Описание архитектуры согласно варианту

Данная модель предназначена для классификации изображений фруктов. На вход модели подаются изображения, выходом является предсказанный класс входного изображения – принадлежность к какому-либо виду фруктов. Представленная модель нейронной сети классификации изображений имеет архитектуру, показанную на рисунке 1. Данная модель имеет несколько слоев:

1) входной слой Rescaling – используется для изменения размера входных изображений;

2) слой Conv2D – блок свертки с 16 фильтрами и ядром свертки размером 3;

3) слой maxpooling2d – слой производит более жесткую свертку изображения, сразу уменьшая размерность каждого слоя, например, в 2 раза (рисунок 2);

4) слой Conv2D – блок свертки с 32 фильтрами и ядром свертки размером 3;

5) слой maxpooling2d;

6) слой Conv2D – блок свертки с 64 фильтрами и ядром свертки размером 3;

7) слой maxpooling2d;

8) слой Flatten – переводит полученный тензор в вектор;

9) слой Flatten;

10) слой Dense – полносвязный слой размерностью 128;

11) выходной слой Dense – полносвязный слой размерностью, которая равна количеству классов.

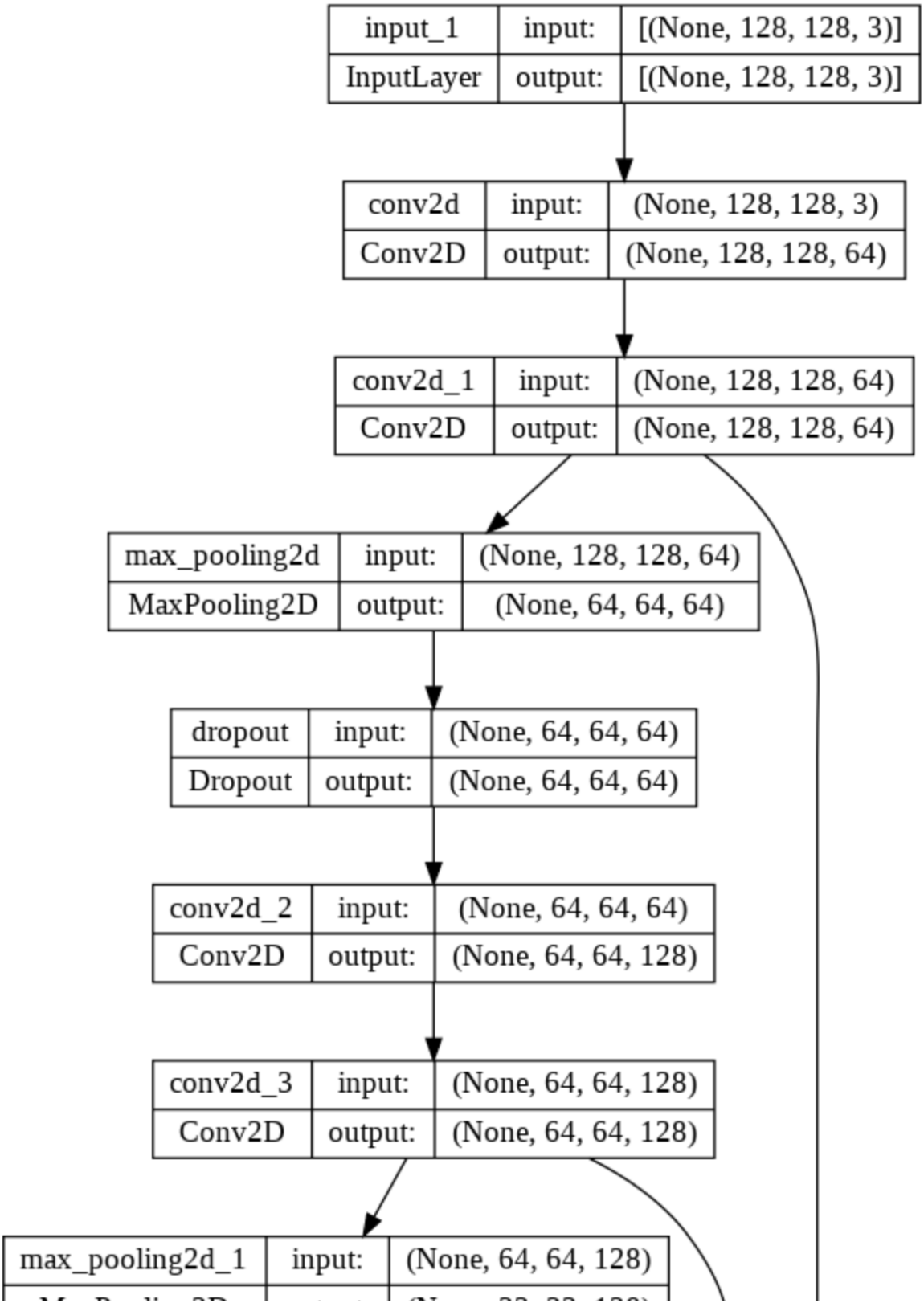


Рисунок 1 – Визуализация архитектуры модели

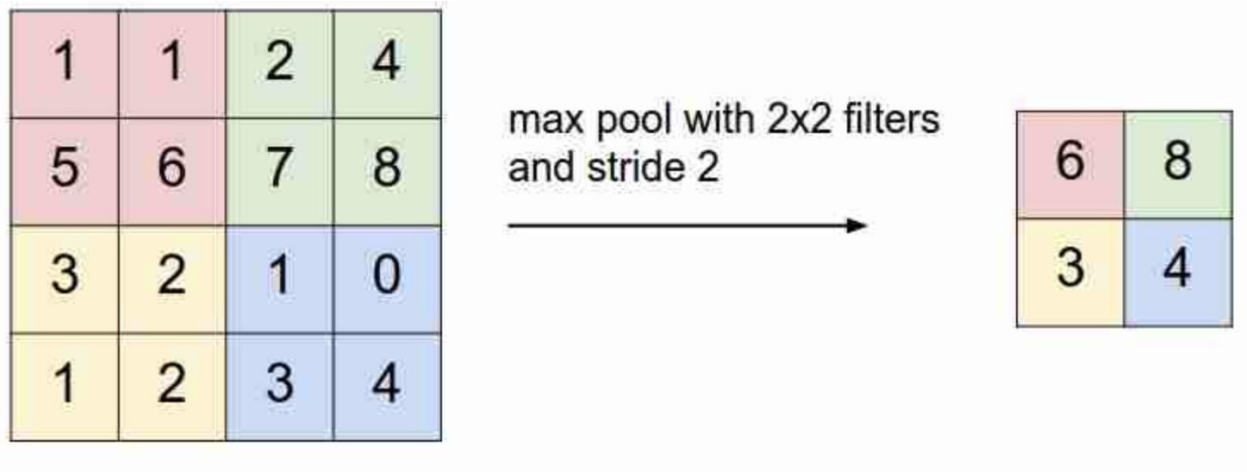


Рисунок 2 – Преобразование MaxPooling

Верификация модели нейронной сети

Для проведения верификации модели нужно загрузить несколько изображений, принадлежащих к различным классам, для проверки работоспособности модели (рисунок 3).

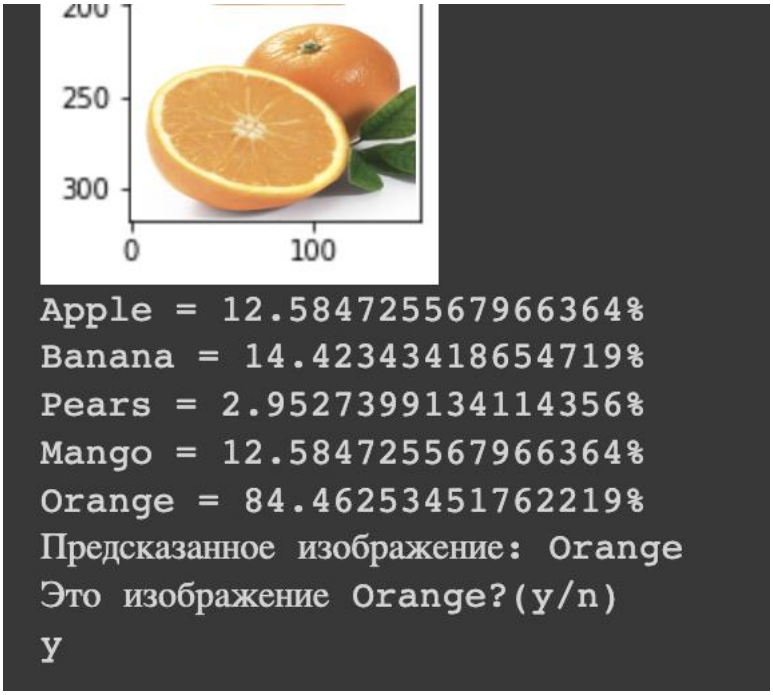


Рисунок 3 – Верификация модели

После тестирования модели на нескольких различных образцах, можно сделать вывод, что модель действительно проводит классификацию фруктов, изображенных на картинках.

Проверка программного кода

Исследуемый программный код рассматриваемой модели имеет несколько недостатков:

1) наличие дублированного слоя Flatten. Так как слой Flatten переводит тензор в вектор, то дублирование этого слоя в данном случае не нужно;

2) неверные пропорции разделения обучающей и валидационной выборки.

Оценка качества обучения и работы модели

Была произведена модификация представленного программного кода с помощью библиотеки для визуализации matplotlib, которая предоставляет инструменты визуализации. График обучения модели представлен на рисунке 4. Как видно, ошибка в процессе проверки модели на валидационной выборке растет, это связано с выше перечисленными недостатками.

Рисунок 4 – График обучения и валидации модели

Выводы

В ходе выполнения работы были описаны и охарактеризованы составные ноды модели нейронной сети классификации изображений, их взаимосвязи и входные/выходные размерности. Также была получена оценка точности представленной модели на тестовом наборе данных с использованием гиперпараметров по умолчанию.

По результатам эксперимента были выявлены оптимальные гиперпараметры модели, имеющие наибольшие показатели точности предсказания модели.

Попов Илья Юрьевич
Бучаев Абдулхамид Яхьяевич
Есипов Дмитрий Андреевич

Валидация систем искусственного интеллекта

Лабораторный практикум

В авторской редакции
Редакционно-издательский отдел Университета ИТМО
Зав. РИО Н.Ф. Гусарова
Подписано к печати
Заказ №
Тираж
Отпечатано на ризографе

Редакционно-издательский отдел
Университета ИТМО
197101, Санкт-Петербург, Кронверкский пр., 49