Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ ИУ «Информатика, искусственный интеллект и системы управления»

КАФЕДРА ИУ-1 «Системы автоматического управления»

ОТЧЕТ ПО ПРОИЗВОДСТВЕННОЙ ПРАКТИКЕ

Студент Юдаков Дмитрий Игоревич
фамилия, имя, отчество

Группа ИУ1-81Б

Тип практики преддипломная

Название предприятия ФГУП «ГосНИИАС»

Студент	28/05/2022 (Подпись, дата)	Д.И. Юдаков (И.О. Фамилия)
Руководитель практики от предприятия	28/05/2022 (Подпись, дата)	В.А. Князь (И.О. Фамилия)
Руководитель практики от МГТУ	28/05/2022 (Подпись, дата)	К.В. Парфентьев (И.О. Фамилия)

Оценка _____

2022 г.

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

УТВЕРЖДАЮ

Заведующий кафедрой ИУ-1
(Индекс)

_____ К.А. Неусыпин
(И.О. Фамилия)

« 15 » мая 20 22 г.

З А Д А Н И Е
на прохождение производственной практики

Студент группы ИУ1-81Б
Юдаков Дмитрий Игоревич
(Фамилия, имя, отчество)

Задание
Найти и описать существующие наборы данных с жестами кисти.
Дообучить нейронную сеть классификации из выпускной квалификационной работы на найденных наборах данных.
Провести сравнительный анализ результатов до дообучения и после.

Оформление отчета по практике:
Отчет на 18 листах формата А4.
Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)
оформление графического материала в отчете по практике не предусмотрено

Дата выдачи задания « 15 » мая 20 22 г.

Руководитель Практики от предприятия	15/05/2022 (Подпись, дата)	В.А. Князь (И.О. Фамилия)
Руководитель Практики от МГТУ	15/05/2022 (Подпись, дата)	К.В. Парфентьев (И.О. Фамилия)
Студент	15/05/2022 (Подпись, дата)	Д.И. Юдаков (И.О. Фамилия)

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Вычислительная способность компьютера стала необходимой в жизни современного человека – без неё довольно трудно жить и работать в нынешних реалиях. Умение быстро помочь в решении самых сложных задач начиная от проведения необходимых расчётов до оптимизации управления предприятием высоко оценивается человечеством. Камнем преткновения остаётся интерфейс взаимодействия системы «человек-машина» – оно должно быть максимально упрощено, чтобы люди и машины коммуницировали на уровне естественных средств общения.

Развитие компьютерных технологий получило огромный прирост за последнюю половину века и стремительно увеличивается и по сей день. Они уже способны обрабатывать и анализировать информацию подобно человеку: распознавать текст, изображения, анализировать звуки и мелодии, произносить осмысленные предложения, распознавать голосовые команды и реагировать на прикосновения пальцев. Но одной из самых приоритетных и трудных задач в области информационных технологий и интеллектуальных систем является задача распознавания жестов.

Актуальность рассматриваемой тематики обусловлена возможностью применения предлагаемого подхода для управления объектов без тактильного контакта и голосовой идентификации команд, а также своей простотой с точки зрения конечного пользователя.

Во время прохождения практики будет произведён поиск существующих наборов данных с жестами кисти, дообучение сформированной в выпускной квалификационной работе нейронной сети классификации жестов и сравнение качества работы до и после дообучения.

1. МЕТОДИКА КЛАССИФИКАЦИИ ЖЕСТОВ

Классификация жестов происходит на основе прогона ключевых точек, полученных с изображения кисти через искусственную нейронную сеть с архитектурой, представленной на рис. 1.

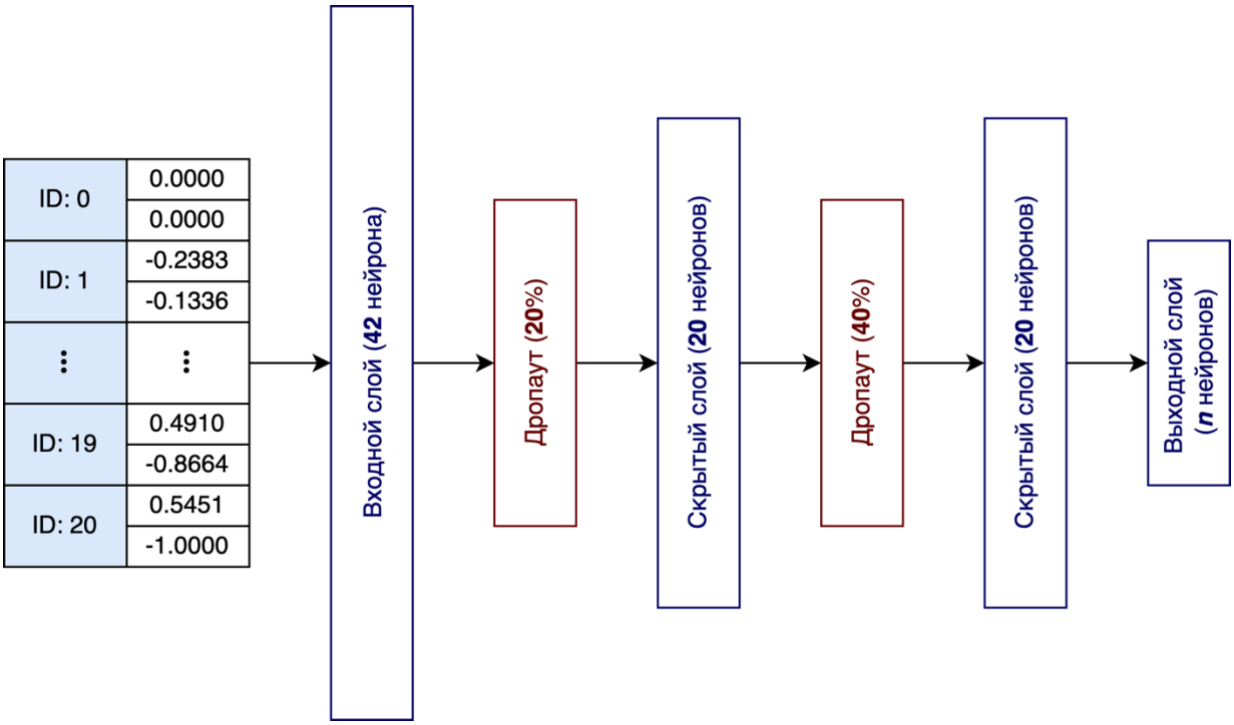


Рис. 1. Архитектура нейронной сети для классификации жестов.

Более подробно структуру можно записать следующим образом:

Входной слой (42×1)

Дропаут 20%

Скрытый слой (20×1) с функцией активации ReLU:

$$f(x) = \begin{cases} 0, x < 0 \\ x, x \geq 0 \end{cases}$$

Дропаут 40%

Скрытый слой (10×1) с функцией активации ReLU

Выходной слой $(n \times 1)$ с функцией активации Softmax, где $n -$ количество жестов.

Исключение или *дропаут* – метод регуляризации искусственных нейронных сетей, предназначен для уменьшения переобучения сети за счёт предотвращения сложных коадаптаций отдельных нейронов на тренировочных данных во время обучения, характеризует исключение определённого процента случайных нейронов на разных итерациях во время обучения нейронной сети. Такой приём значительно увеличивает скорость обучения,

качество обучения на тренировочных данных, а также повышает качество предсказаний модели на новых тестовых данных.

Более подробно рассмотрим способ поиска ключевых точек.

1.1. Поиск ключевых точек на основе MediaPipe Hands

Одним из применений Single Shot детектора (SSD) в задаче поиске кисти на изображении является подход, реализованный в фреймворке MediaPipe от Google – MediaPipe Hands [1]. Для определения начального положения рук была разработка модель SSD, оптимизированная для использования в реальном времени. Поскольку обнаружение рук чрезвычайно сложная задача из-за огромного количества их конфигураций, то вместо детектора рук обучался детектор ладоней, так как получать прямоугольники негнущихся объектов, таких как ладони и кулаки, значительно проще, нежели обнаружение рук вместе с пальцами. Кроме того, поскольку ладони являются более мелкими объектами, алгоритм подавления немаксимумов хорошо работает даже для случаев пересечения двух рук, таких как рукопожатия. Более того, ладони могут быть смоделированы с помощью квадратных анкеров, игнорируя другие соотношения сторон, уменьшая количество анкеров в 3-5 раз. Во-вторых, кодирующий-декодирующий экстрактор признаков используется для большей осведомленности о контексте сцены даже для небольших объектов. Наконец, во время обучения минимизируется функция кросс-энтропийных потерь с динамическим масштабированием (focal loss), чтобы поддерживать большое количество анкеров, возникающих из-за большой дисперсии масштаба. С помощью описанных выше методов достигается средняя точность обнаружения ладони 95,7%. Использование обычной функции кросс-энтропийных потерь и без декодера даёт точность всего 86,22%.

После обнаружения ладони на всём изображении последующая модель выполняет точную локализацию 21 ключевой точки (рис. 2) внутри обнаруженных областей руки посредством регрессии (как для определения цифры, написанной рукой), то есть прямого прогнозирования координат. Модель обучается последовательному внутреннему представлению позиции рук и устойчива даже к частично видимым рукам и самопересечениям. Так же стоит отметить, что кроме 21 ключевой точки модель так же определяет вероятность присутствия кисти на кадре и классификацию типа кисти – левой или правой.

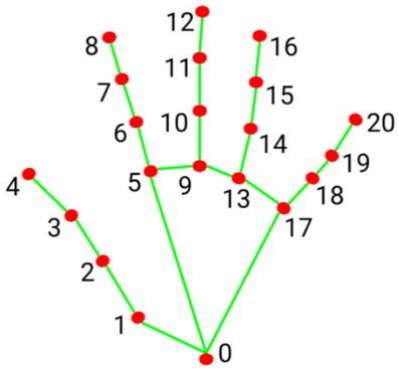


Рис. 2. Ключевые точки на руке, определяемые с помощью MediaPipe Hands.

Чтобы получить данные для обучения, вручную было промаркировано около 30 000 изображений реального мира с 21 координатой, как показано на рис. 3. А также, чтобы лучше охватить возможные позиции рук и обеспечить дополнительное наблюдение за характером геометрии рук, была визуализирована высококачественная синтетическая модель руки на различных фонах, таким образом синтезируя набор из 100 000 изображений. Для детектора ладони использовались лишь 6 000 изображений данного набора, поскольку этого достаточно для его обучения локализации кисти.

Таким образом, библиотека состоит из детектора ладоней, определяющего область изображения, на которой изображена рука и модели трекинга руки, предсказывающая ключевые точки на области, полученной детектором.

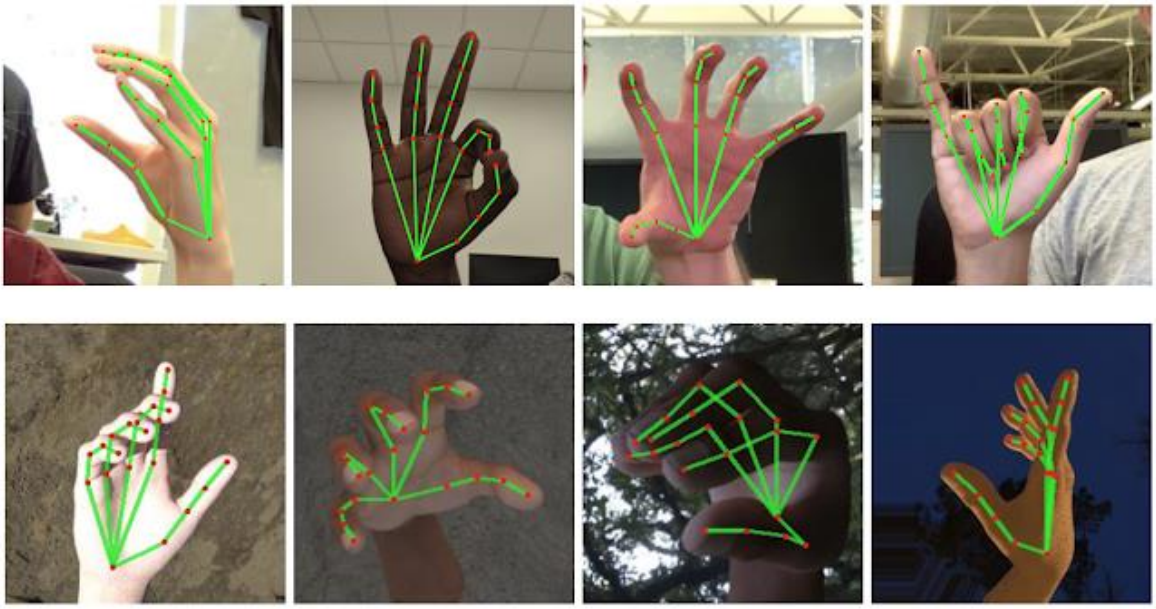


Рис. 3. Сверху: реальные изображения рук, аннотированные вручную.
Снизу: сгенерированные синтетические изображения рук с аннотациями.

Основываясь на данной технологии, полученные ключевые точки можно использовать в дальнейших приложениях, одним из которых является классификация жестов.

1.2. Генерация набора данных и обучение нейросети

Для быстрого и эргономичного наполнения обучающей выборки данными был спроектирован графический пользовательский интерфейс, позволяющий контролировать количество уже записанных жестов, а также имеющий режим непрерывной записи жеста.

Спроектированный вид интерфейса изображён на рис. 4. В верхней части выводится количество записей по каждому из жестов, предоставляя возможность выравнивать количество данных по ним. При нажатии на клавишу

- *‘q’* или *Escape* – программа завершает свою работу;
- *Цифру* – начинается непрерывная запись или записывается жест под нажатой цифрой;
- *‘c’* – переключаются режимы одиночной и непрерывной записи жеста. В первом режиме координаты ключевых точек записываются только при нажатии на соответствующую цифру. Последний же позволяет сохранять координаты ключевых точек непрерывно, а не только при нажатии на цифру. Эта способность позволяет намного быстрее собирать большое количество данных, в процессе немного модифицируя и перемещая руку с жестом.

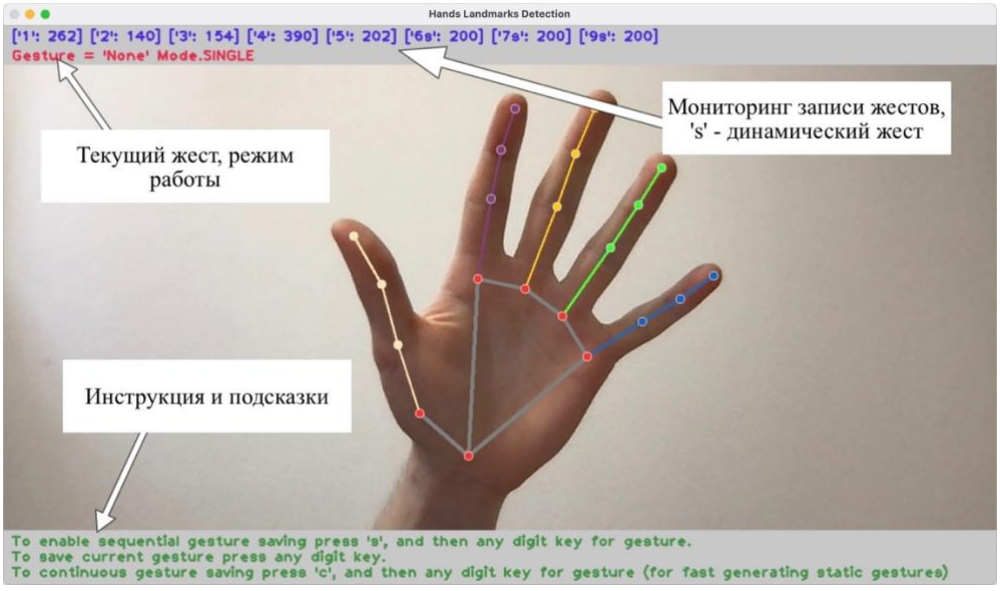
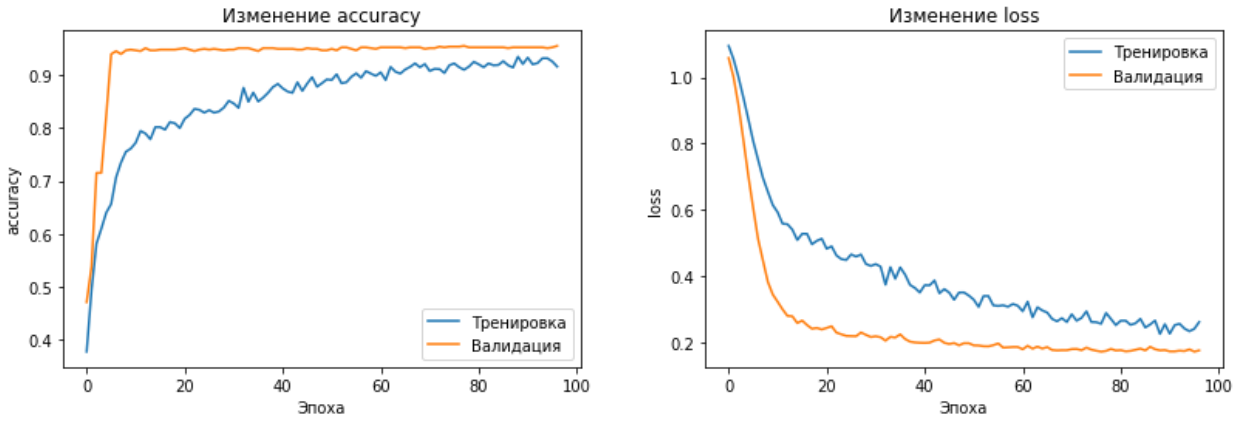


Рис. 4. Вид программы.

Таким образом был собран набор из 3 жестов с вытянутым одним, двумя или тремя пальцами соответственно, состоящий из 3138 записей наборов ключевых точек.

Обучение для собранного набора проходило на 97 *эпохах (рис. 5). Доля правильных ответов (accuracy) составила* 0,9656*, а потери (loss)* – 0,1423*. Основные метрики по каждому классу представлены в таблице*

.



а) б)

Рис. 5. Изменение метрики accuracy (а) и loss (б) на тренировочных и валидирующих данных во время обучения.

Таблица 1.

Показатели метрик классификации.

Класс	Точность (precision)	Полнота (recall)	F-мера (f1-score)
1	0,96	1,00	0,98
2	0,95	0,97	0,96
3	1,00	0,91	0,95

Рассмотрим каждую из метрик. Перед переходом к самим метрикам необходимо ввести важную концепцию для описания метрик в терминах ошибок классификации – confusion matrix (матрица ошибок). Допустим, что существует два класса и алгоритм, предсказывающий принадлежность каждого объекта к одному из классов, тогда матрица ошибок классификации будет выглядеть так, как представлено в таблице 2.

Таблица 2.

Матрица ошибок для бинарной классификации.

	$y = 1$	$y = 0$
$\hat{y} = 1$	True Positive (TP)	False Positive (FP)
$\hat{y} = 0$	False Negative (FN)	True Negative (TN)

Здесь $\hat{y}$ – это ответ алгоритма на объекте, а y – истинная метка класса на этом

объекте. Таким образом, ошибки классификации бывают двух видов: False Negative (FN) и False Positive (FP). Матрица ошибок для задачи распознавания жестов представлена на рис. 6.

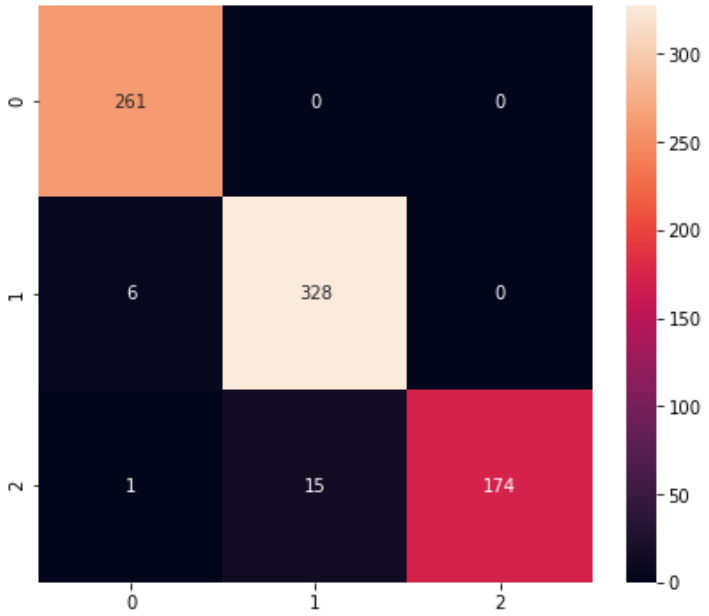


Рис. 6. Матрица ошибок для задачи распознавания жестов.

Интуитивно понятной, очевидной и почти неиспользуемой метрикой является accuracy – доля правильных ответов алгоритма:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Для оценки качества работы алгоритма на каждом из классов по отдельности вводятся метрики precision (точность) и recall (полнота):

$$\text{precision} = \frac{TP}{TP + FP}, \text{recall} = \frac{TP}{TP + FN}.$$

Precision можно интерпретировать как долю объектов, названных классификатором положительными и при этом действительно являющимися положительными, а recall показывает, какую долю объектов положительного класса из всех объектов положительного класса нашёл алгоритм. Существует несколько различных способов объединить precision и recall в агрегированный критерий качества. F-мера (в общем случае F_β) – среднее гармоническое precision и recall:

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}.$$

Переменная β в данном случае определяет вес точности в метрике, и при $\beta = 1$ это среднее гармоническое. F-мера достигает максимума при полноте и точности, равными единице, и близка к нулю, если один из аргументов близок к нулю.

Проанализировав результаты, можно сделать вывод, что модель успешно обучена и её можно применять в работе. Но при использовании сети на большем количестве жестов метрики резко падают, а значит архитектура сети нуждается в улучшении.

2. УЛУЧШЕННАЯ МЕТОДИКА РАСПОЗНАВАНИЯ ЖЕСТОВ

Улучшение методики распознавания жестов заключается в проектировании новой архитектуры нейросети, способной качественно работать при большом количестве жестов. Для оптимизации архитектуры следует произвести поиск больших уже собранных наборов данных с большим количеством разнообразных жестов.

2.1. Поиск наборов данных с жестами

В результате поиска открытых наборов данных были отобраны три источника, первый из них это набор HANDS статичных жестов для взаимодействия человека и робота [1]. Этот набор был создан для исследования взаимодействия между человеком и роботом и состоит из пространственно и временно выровненных кадров RGB и глубины. Он содержит 12 видов статических жестов, выполненных как правой, так и левой рукой, и 3 вида статических жестов с двумя руками, таким образом, всего 26 уникальных классов.

Пять испытуемых (2 женщины и 3 мужчины) выполняли жесты, каждый из них на разном фоне и в разных условиях освещения. Для каждого жеста было собрано 150 кадров RGB и соответствующие им 150 кадров глубины, всего 2400 кадров RGB и 2400 кадров глубины на человека.

Данные были собраны с помощью камеры Kinect V2, откалиброванной для пространственного выравнивания данных RGB с данными о глубине. Временное выравнивание было выполнено в автономном режиме с использованием MATLAB, выравнивая кадры с максимальным временным расстоянием 66 мс. Примеры жестов изображены на рис. 7.

Рис. 7. Примеры фотографий каждого класса из набора данных HANDS.

Второй – набор статичных жестов, собранный с помощью технологий Leap Motion и Kinect [2, 3]. Этот набор содержит 10 видов жестов, выполненных 14 разными людьми по 10 раз, таким образом, всего 1400 фотографий. Примеры жестов изображены на рис. 8.

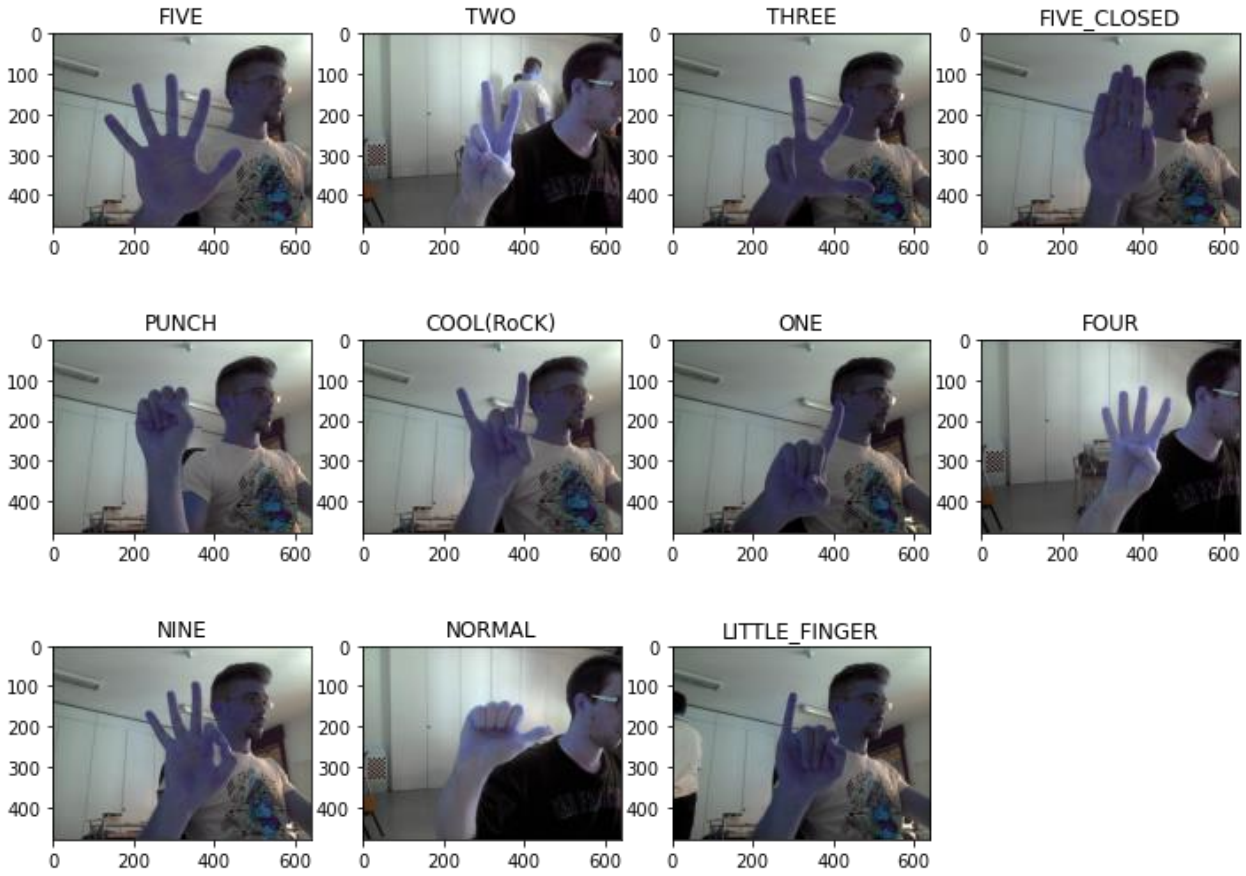


Рис. 8. Примеры жестов из набора данных, собранных с использованием технологии Leap Motion.

Третий – набор из открытого задания на распознавание жестов [5] с сайта Kaggle, состоящий из 10 разных жестов, выполненными 10 разными людьми. Примеры жестов изображены на рис. 9.

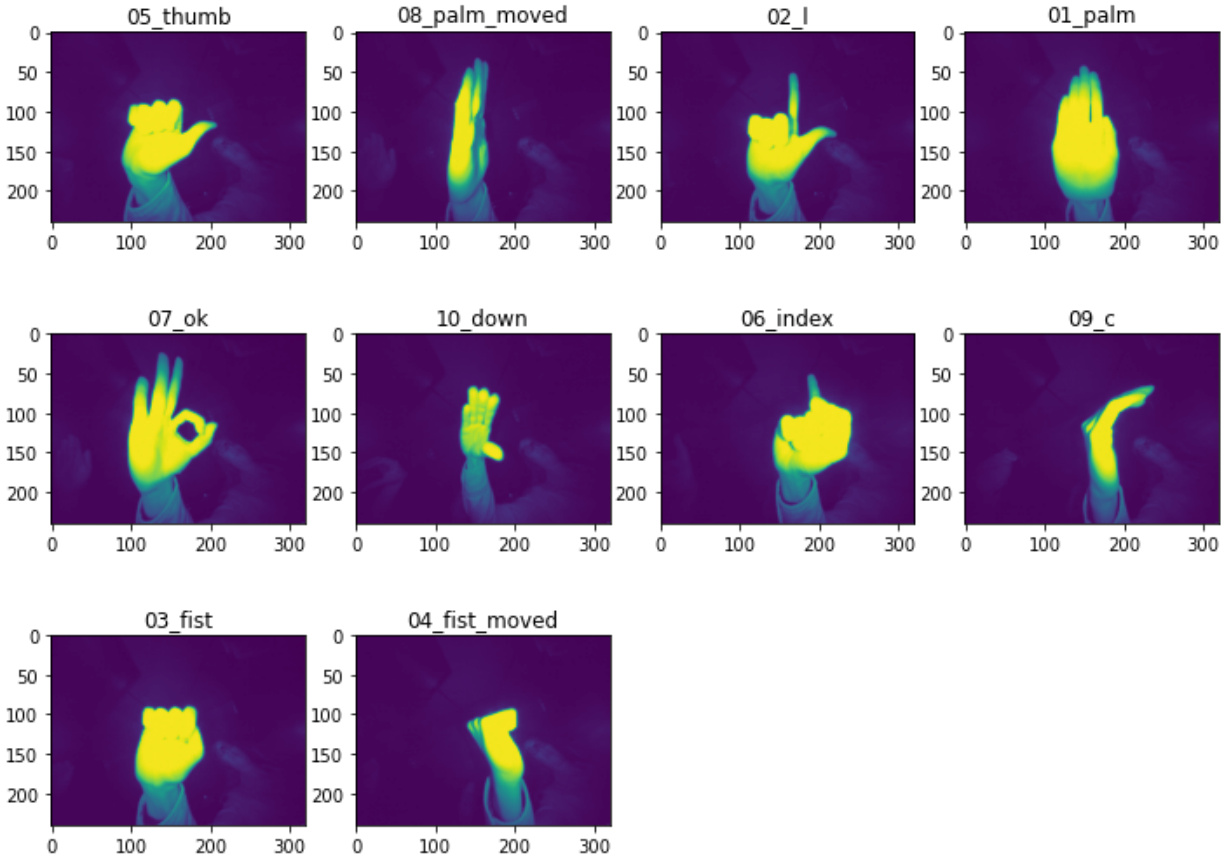


Рис. 9. Примеры жестов из набора данных с сайта Kaggle.

Таким образом, почти на каждом изображении из набора данных были найдены ключевые точки кисти с помощью MediaPipe Hands и общий набор составил 23115 записей, с 15704, 1229 и 6182 записями из каждого набора соответственно. Поскольку первый набор самый объёмный и по количеству разных жестов, и по количеству записей на каждый жест, то именно его классы будут браться за основу и дополняться записями из остальных таких же или похожих жестов.

2.2. Разработка улучшенной архитектуры нейросети и её обучение

Поскольку все основные метрики при обучении на нейросети с прежней архитектурой не превышают значение 0,6, то было решено её модифицировать. Улучшенная структура искусственной нейронной сети для классификации жестов содержит большее количество скрытых слоёв и изображена на рис. 10. Все слои, кроме последнего, как и прежде имеют функцию активации ReLU.

Обучение сети проходило на 156 эпохах (рис. 11). Проверив работу нейросети на тестовых данных, доля правильных ответов (accuracy) составила 0,9924, а потери (loss) – 0,0288. Основные метрики по каждому классу представлены в таблице 3.

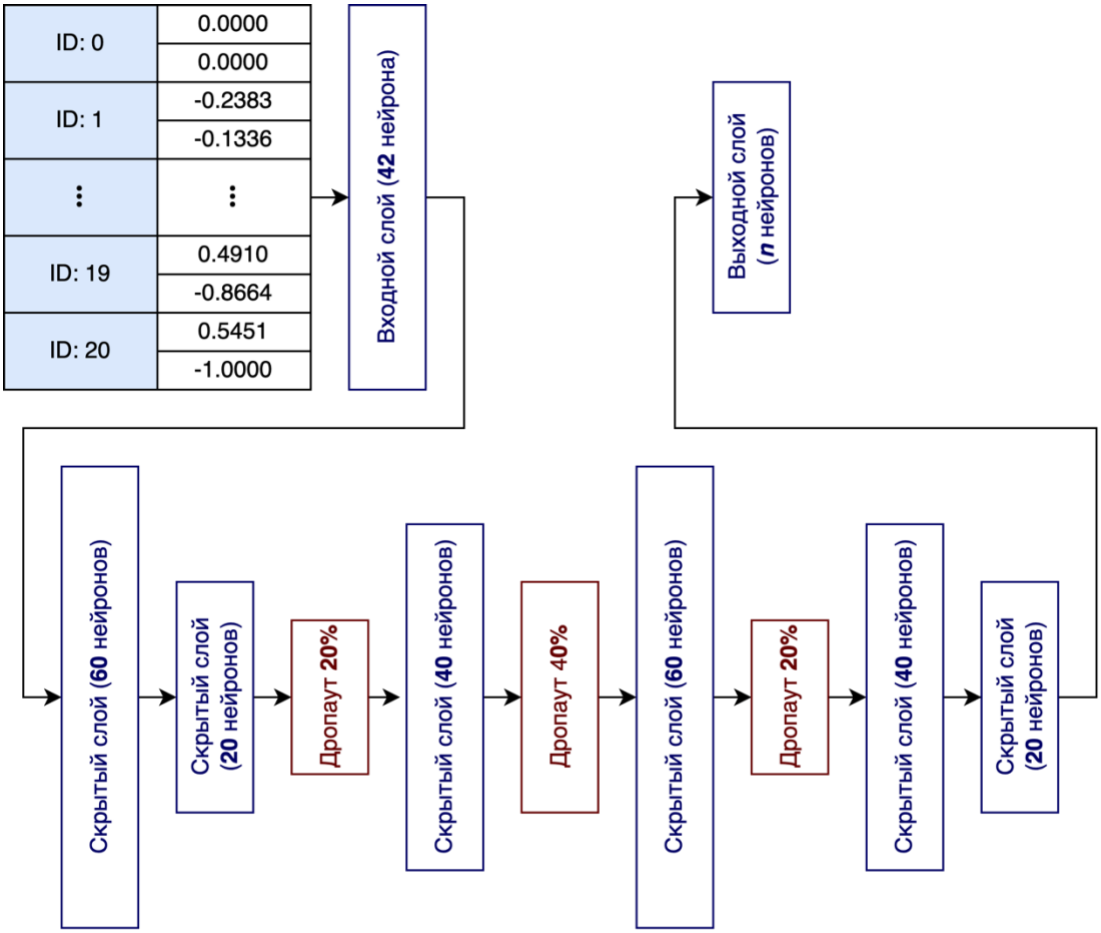
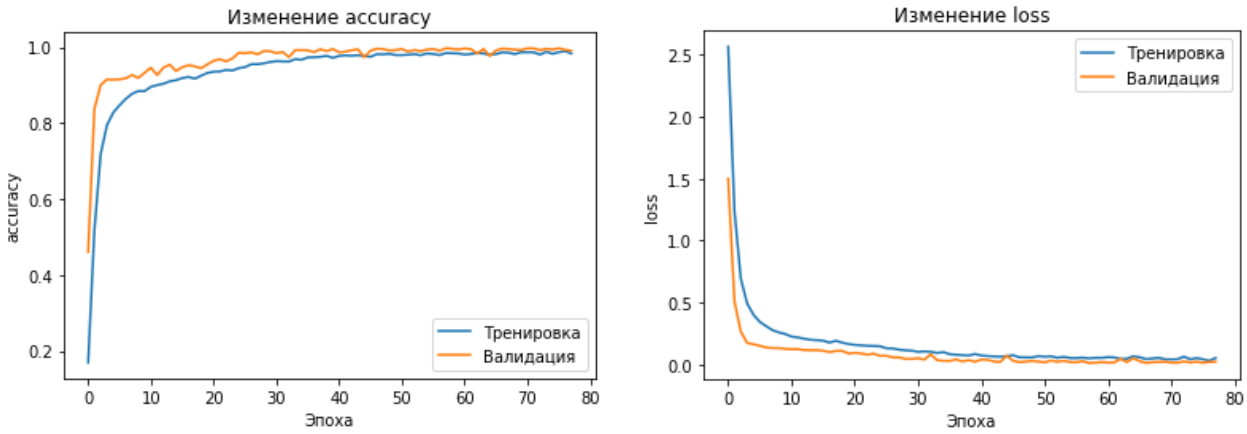


Рис. 10. Улучшенная архитектура нейросети.



а) б)

Рис. 11. Изменение метрики accuracy (а) и loss (б) на тренировочных и валидирующих данных во время обучения улучшенной нейросети.

Таблица 3.

Показатели метрик классификации.

Класс	Точность (precision)	Полнота (recall)	F-мера (f1-score)
0	1,00	1,00	1,00
1	0,99	0,99	0,99
2	1,00	1,00	1,00
3	1,00	1,00	1,00
4	1,00	1,00	1,00
5	1,00	1,00	1,00
6	1,00	1,00	1,00
7	1,00	1,00	1,00
8	1,00	1,00	1,00
9	0,99	1,00	1,00
10	0,99	1,00	1,00
11	1,00	1,00	1,00
12	1,00	0,99	1,00
13	0,99	1,00	1,00
14	1,00	1,00	1,00
15	1,00	1,00	1,00
16	1,00	1,00	1,00
17	0,99	1,00	1,00
18	0,99	1,00	1,00
19	0,99	1,00	0,99
20	1,00	0,98	0,99
21	0,98	1,00	0,99
22	1,00	0,97	0,99
23	1,00	0,95	0,97
24	0,88	1,00	0,93
25	1,00	0,89	0,94

Матрица ошибок для задачи распознавания жестов с помощью улучшенной нейросети представлена на рис. 12.

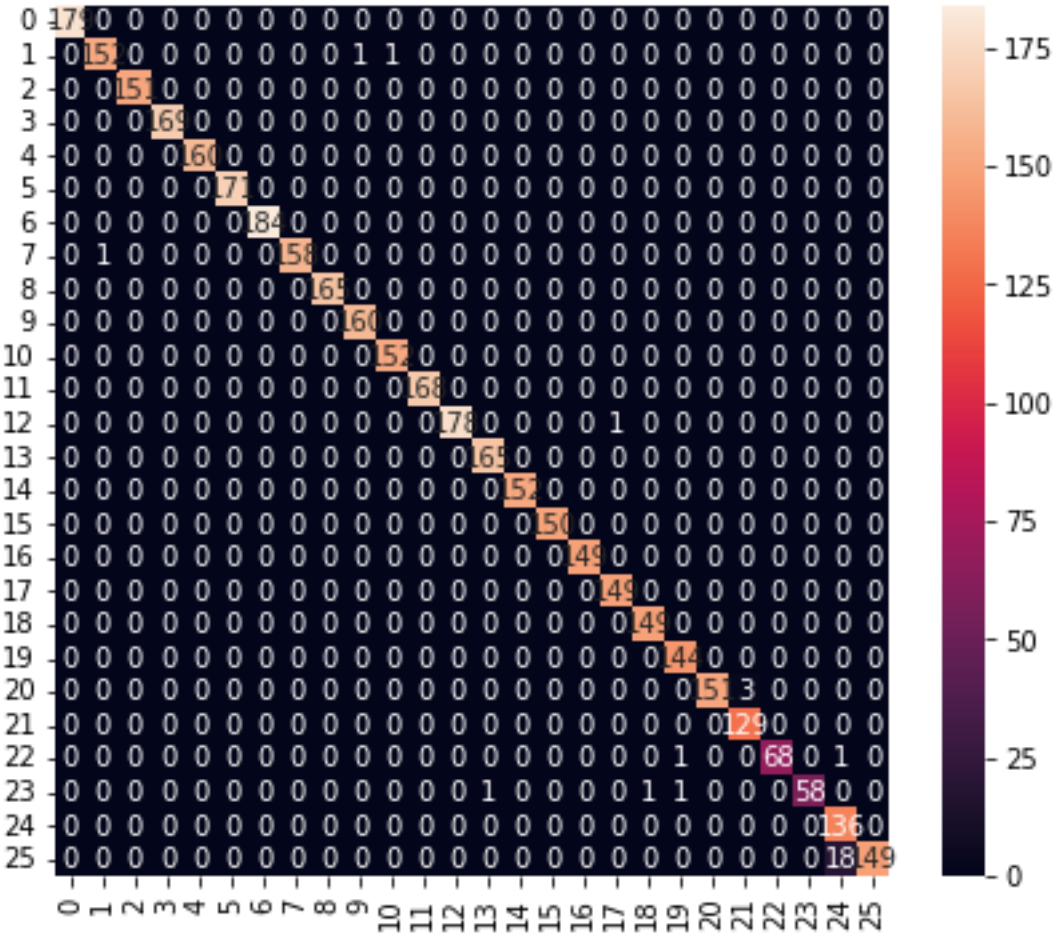


Рис. 12. Матрица ошибок для задачи распознавания жестов с помощью улучшенной нейросети.

Проанализировав результаты, можно сделать вывод, что нейросеть успешно обучена и её можно применять в задаче классификации 26 разных типов жестов.

ЗАКЛЮЧЕНИЕ

В процессе прохождения практики было найдено три уже собранных и опубликованных в открытом доступе набора данных с жестами, а также было проведено их описание и подготовка к обучению.

В процессе обучения на большом наборе данных было выявлено, что первоначальная искусственная нейронная сеть не справляется с большим количеством классов жестов из-за простоты её архитектуры, для этого стало необходимо добавить в неё несколько дополнительных скрытых слоёв и увеличить в них количество нейронов. Таким образом архитектура нейросети расширилась до 6 скрытых слоёв с 3 слоями дропаута.

По итогу работы на тестовых данных основные метрики работы улучшенной нейросети, а именно accuracy и loss, составили $0{,}9924$ и $0{,}0288$ соответственно, что является довольно неплохим результатом.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Fan Z., Valentin B., Andrey V., Andrei T., George S., Chuo-Ling C., Matthias G. MediaPipe Hands: On-device Real-time Hand Tracking. CVPR, 2020.

2. G. Marin, F. Dominio, P. Zanuttigh, "Hand gesture recognition with Leap Motion and Kinect devices", IEEE International Conference on Image Processing (ICIP), Paris, France, 2014.

3. G. Marin, F. Dominio, P. Zanuttigh, "Hand Gesture Recognition with Jointly Calibrated Leap Motion and Depth Sensor", Multimedia Tools and Applications, 2015.

4. Nuzzi Cristina; Pasinetti Simone; Pagani Roberto; Coffetti Gabriele; Sansoni Giovanna (2021), “HANDS: a dataset of static Hand-Gestures for Human-Robot Interaction”, Mendeley Data, V2, doi: 10.17632/ndrczc35bt.2.

5. T. Mantecón, C.R. del Blanco, F. Jaureguizar, N. García, “Hand Gesture Recognition using Infrared Imagery Provided by Leap Motion Controller”, Int. Conf. on Advanced Concepts for Intelligent Vision Systems, ACIVS 2016, Lecce, Italy, pp. 47-57, 24-27 Oct. 2016. (doi: 10.1007/978-3-319-48680-2_5).