Ву Минь Куанг, студент
МГТУ им. Н. Э. Баумана

Быстрицкая Анна Юрьевна, к.т.н.
МГТУ им. Н. Э. Баумана

ИССЛЕДОВАНИЕ МЕТОДА ПЕРЕВОДА ЖЕСТОВ РУК В ТЕКСТ С ИСПОЛЬЗОВАНИЕМ РЕКУРРЕНТНЫХ НЕЙРОННЫХ СЕТЕЙ

Аннотация. В данной работе представлен метод перевода жестов рук в текст с использованием рекуррентных нейронных сетей LSTM и библиотеки MediaPipe для отслеживания движений рук. Проведена комплексная оценка качества распознавания жестов, исследовано влияние настроечных параметров модели на точность классификации и скорость работы системы

Ключевые слова: Распознавание жестов, LSTM, MediaPipe, компьютерное зрение

Введение

Жестовый язык является основным средством коммуникации для более чем 70 миллионов глухих людей по всему миру. Однако существует значительный коммуникационный барьер между людьми, использующими жестовый язык, и теми, кто его не знает. Разработка автоматизированных систем распознавания жестового языка является актуальной задачей, решение которой может значительно улучшить качество жизни людей с нарушениями слуха [1].

Современные подходы к распознаванию жестов используют методы глубокого обучения, в частности, рекуррентные нейронные сети (RNN) и их модификации, такие как LSTM (Long Short-Term Memory) [2, 3]. Преимуществом LSTM-сетей является их способность эффективно работать с последовательностями данных и учитывать временные зависимости, что критично для распознавания динамических жестов.

Архитектура системы

Разработанная система состоит из четырех основных модулей:

Модуль пользовательского интерфейса обеспечивает взаимодействие пользователя с системой распознавания жестов, предоставляя визуальное отображение процесса распознавания и результатов работы системы.

Модель LSTM играет ключевую роль в архитектуре систем распознавания и перевода дактильных жестов в текст на основе глубокого обучения, особенно когда требуется обработка последовательностей данных.

Модуль отслеживания рук и извлечения признаков обеспечивает детекцию и отслеживание ключевых точек тела, лица и рук человека, формируя векторное представление жеста для классификации. Захват видеопотока с камеры и извлечение ключевых точек с использованием MediaPipe Holistic. Извлекаются координаты 126 признаков: 21 точка для каждой руки (×3 координаты).

Модуль перевода жестов рук в текст отвечает за преобразование последовательности жестов в текстовое представление с применением постобработки и формированием грамматически корректных предложений.

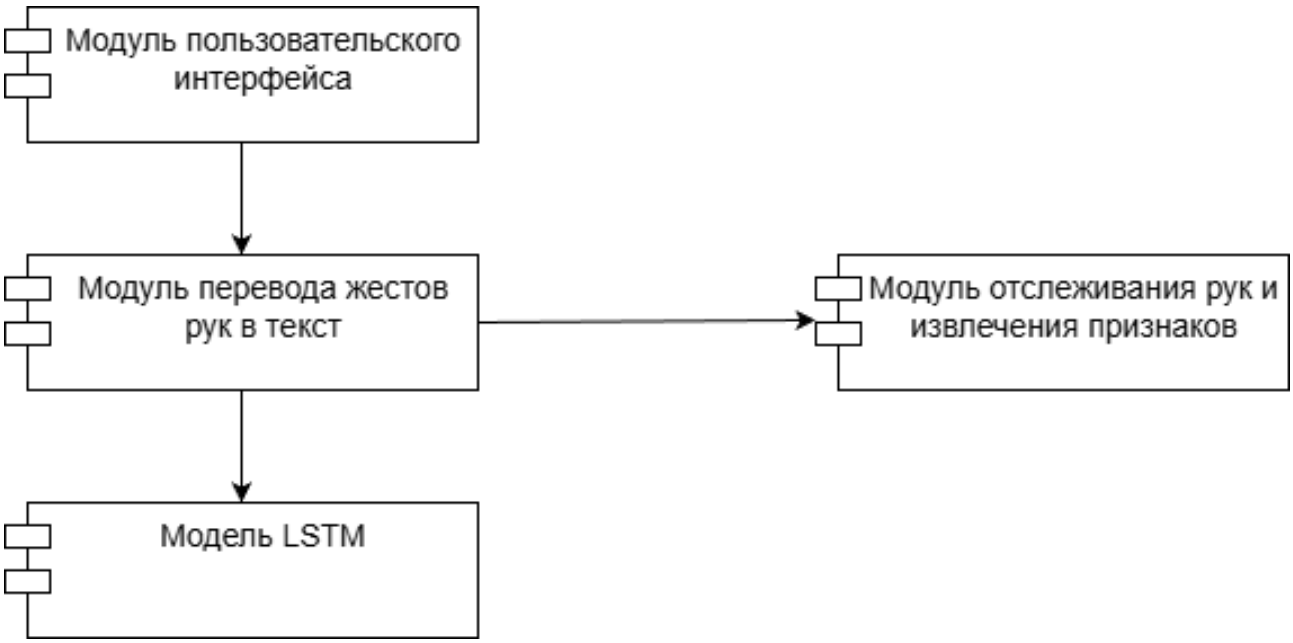

Рисунок 1 – Архитектурная схема разработанного программного обеспечения

Архитектура нейронной сети

Разработанная модель представляет собой последовательную сеть (Sequential) следующей архитектуры:

Слой	**Параметры**	**Выходная размерность**
Первый рекуррентный слой (LSTM-1)	64 юнита, возвращающий последовательности, с функцией активации «ReLU» (64 units, return_sequences=True, ReLU)	(20, 64)
Второй рекуррентный слой (LSTM-2)	128 юнитов, возвращающий последовательности, с функцией активации «ReLU» (128 units, return_sequences=True, ReLU)	(20, 128)
Третий рекуррентный слой (LSTM-3)	64 юнита, возвращает только последний выход, с функцией активации «ReLU» (64 units, return_sequences=False, ReLU)	(64)
Первый полносвязный слой (Dense-1)	64 юнита с функцией активации «ReLU» (64 units, ReLU)	(64)
Второй полносвязный слой (Dense-2)	32 юнита с функцией активации «ReLU» (32 units, ReLU)	(32)
Третий полносвязный слой (Output)	3 класса для классификации (3 units, Softmax)	(3)

Входные данные имеют размерность (20, 126), где 20 – количество кадров в последовательности, 126 – количество признаков. Модель обучалась с использованием оптимизатора Adam и функции потерь «categorical_crossentropy» на протяжении 100 эпох.

Общая схема работы процедуры распознавания дактильных жестов

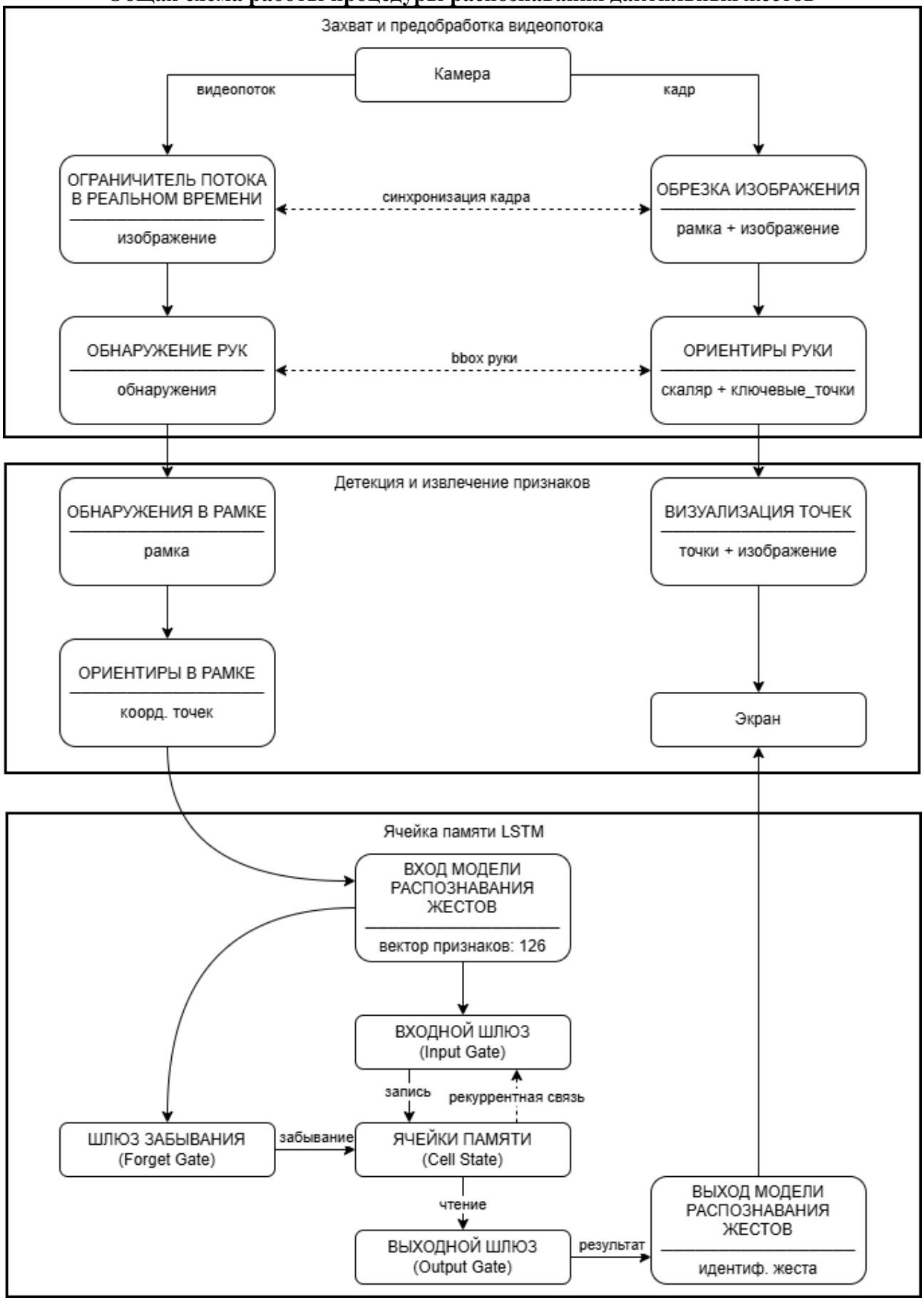

Рисунок 2 – Схема работы процедуры распознавания дактильных жестов

Набор данных

После того, как сбор данных завершен, пользователь может приступить к обучению модели. На этом этапе набор данных разделяется на два подмножества: 90% набора данных используется для обучения и 10% для тестирования. Точность тестирования с использованием этих 10% набора данных даст представление об эффективности модели.

Для этого конкретного проекта нейронная сеть построена с использованием экземпляра последовательной модели путем прохождения трех LSTM и трех плотно связанных слоев. Первые пять из этих слоев используют функцию активации «ReLU», а последний слой – функцию активации «Softmax». В процессе обучения алгоритм оптимизации Adam используется для получения оптимальных параметров для каждого слоя.

Результаты экспериментов

При обучении нейронной сети подбиралось и анализировалось количество эпох. Графики потерь и категориальной точности для обучающейся модели, построенные для 160 эпох, показали переобучение после 130-й эпохи (рис. 3, 4).

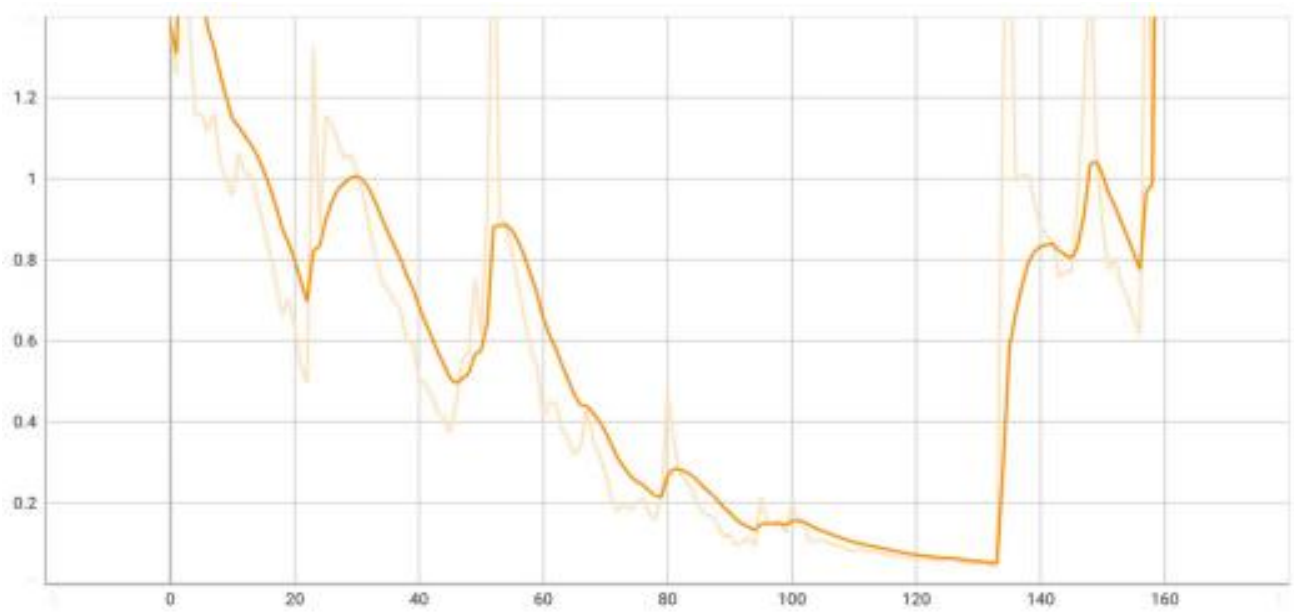

Рисунок 3 – График потерь для 160 эпох.

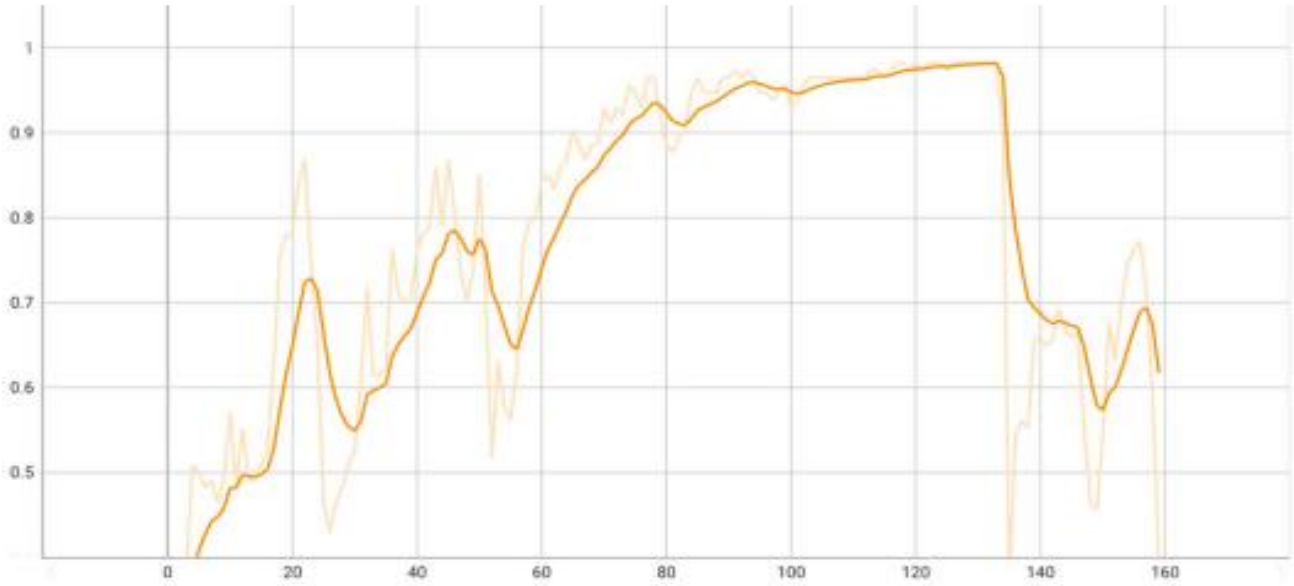

Рисунок 4 – График категориальной точности для 160 эпох.

Для повышения качества распознавания принято решение снизить количество эпох до 120 (рис. 5, 6).

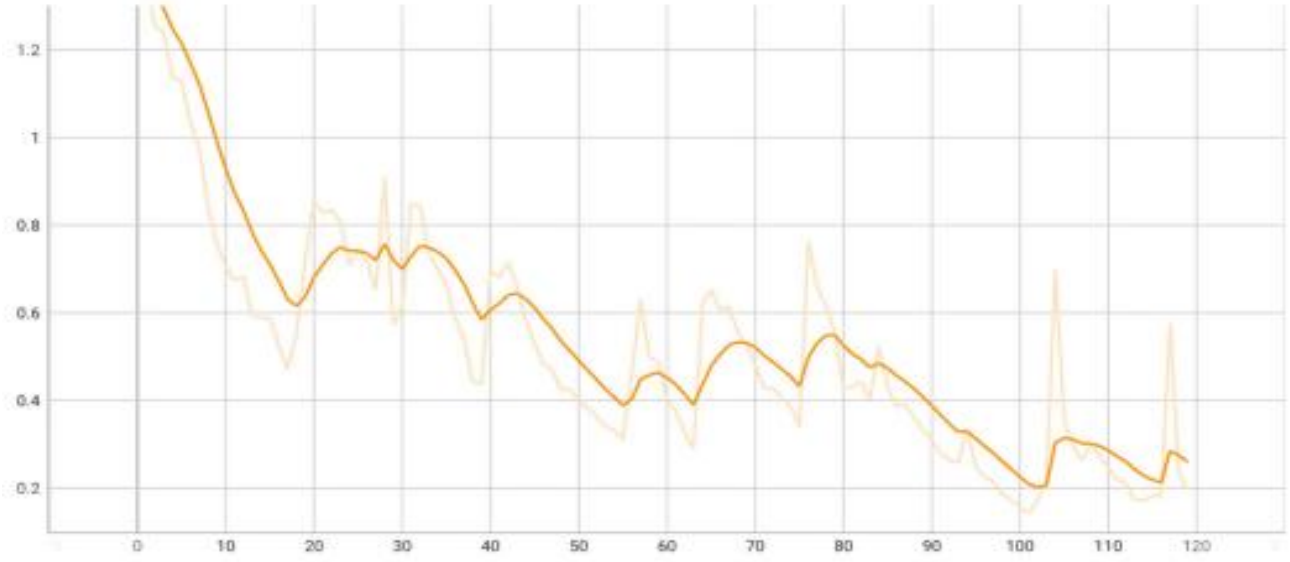

Рисунок 5 – График потерь для 120 эпох.

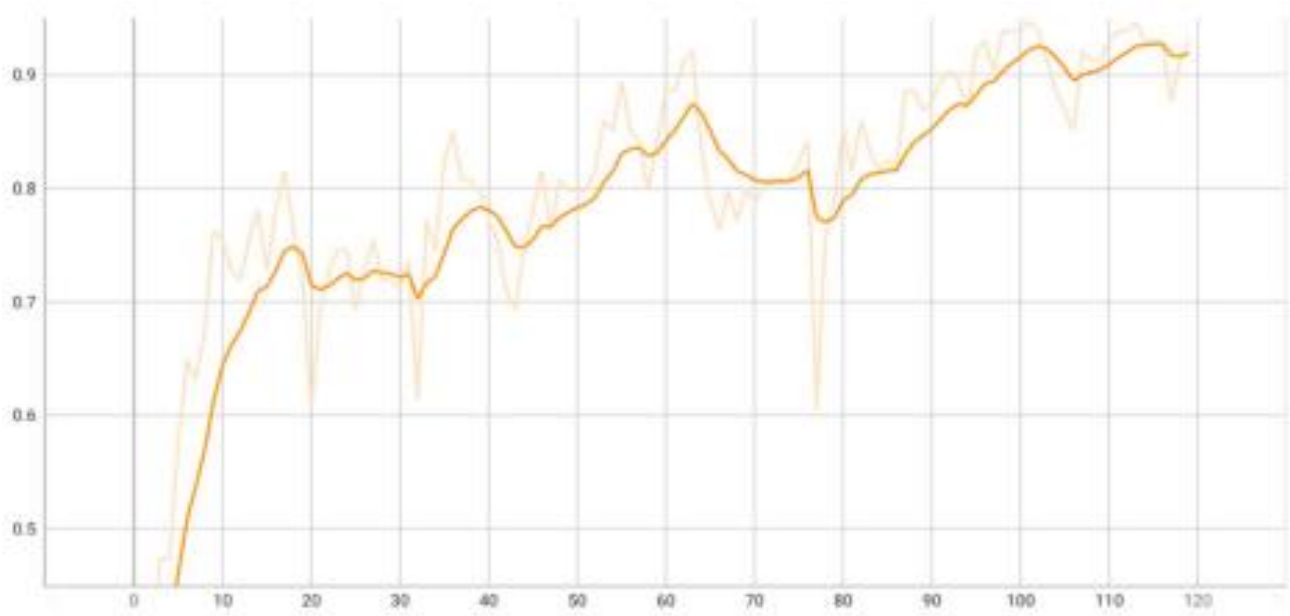

Рисунок 6 – График категориальной точности для 120 эпох.

Результаты показали, что проблема переобучения решилась. При 120 эпохах категориальная точность составляет 92%, а при 160 эпохах – всего 63%, что явно недостаточно для качественного распознавания.

Заключение

Метод перевода дактильных жестов в текст, основанный на совместном использовании рекуррентных нейронных сетей, представляет собой перспективное направление в области компьютерного зрения и обработки естественных языков для людей с ограниченными возможностями слуха. Данный подход успешно решает ключевые задачи: вариабельность исполнения жестов и необходимость анализа временных последовательностей. Программное решение было апробировано на алфавите глухонемых и может быть использовано при обучении сурдопереводчиков.

Предложенная архитектура нейронной сети способна распознавать элементы дактильного русского языка с категориальной точностью 92,54% и может служить основой системы автоматического сурдоперевода и других человеко-машинных интерфейсов, позволяющих взаимодействовать с программами посредством жестов

Список литературы:

1. Д.А. Булыгин, Т.Е. Мамонова, Распознавание жестов рук в режиме реального времени // Системы анализа и обработки данных. – №1 (78), С.25-40, 2020.

2. Д.А. Рюмин, Метод автоматического видеоанализа движений рук и распознавания жестов в человеко-машинных интерфейсах // Научно-технический вестник информационных технологий, механики и оптики. – №4, С. 525-531, 2020.

3. Д.А. Рюмин, И.А. Кагиров, А.А. Аксенов, А.А. Карпов, Аналитический обзор моделей и методов автоматического распознавания жестов и жестовых языков // Информационно-управляющие системы. – №6 (115), С.10-20, 2021