**Корпус данных**

1. Цифры

1.1. Цифры русского жестового языка.

Поскольку не существует такого общепризнанного набора данных, то он создан самостоятельно с помощью волонтёров. Участвовало 19 волонтёров, из которых 10 девушек и 9 мужчин в возрасте от 20 до 55 лет. Волонтёры показывали цифры от 1 до 10 с помощью жестов. Инструкции о том, как показывать цифры жестового русского языка были взяты из проекта «Словарь. Русский жестовый язык» (https://surdo.me). Было записано 38 видео и взято одно видео из проекта «Словарь. Русский жестовый язык». Также было снято 2 видео для левой и правой руки для проверки метрик на реальном видео.

Все видео были раскадрованы и отобраны лучшие фото. В итоге, получилось 10 папок с жестами цифр от 1 до 10 по 420 цветных фотографий на каждый жест.

Набор данных был обработан с помощью фреймворка mediapipe с открытым исходным кодом, представленный Google, который помогает создавать мультимодальные конвейеры машинного обучения, и разделён на ещё два корпуса данных.

* + 1. 10 папок по 300 документов в формате ‘.json’ для каждого жеста, которые содержат информацию: сколько рук обнаружено на фото и массив данных длинной 43 элемента. Массив содержит ключевые точки рук в формате x, y и z: 21 ориентир для правой руки, 21 ориентир для левой руки и точка соотношения положений двух рук относительно друг друга. Если на фотографии обнаружена одна рука, то элементы с индексами после 21 будут нулевыми.

1.1.2. 10 папок по 300 цветных фотографий для каждого жеста только обнаруженных рук размером 64 × 64.

1.2. Цифры американского жестового языка.

1.2.1. Данный набор данных найден на kaggle (<https://www.kaggle.com/datasets/rayeed045/american-sign-language-digit-dataset>) – это набор цифр от 0 до 9 американского языка жестов. Содержит 10 папок по 500 цветных изображений рук на каждый жест с чёрным фоном размером 400 × 400.

1.2.2. Набор данных был обработан с помощью фреймворка mediapipe с открытым исходным кодом, представленный Google, который помогает создавать мультимодальные конвейеры машинного обучения. Был создан ещё один корпус данных, который содержит 10 папок по 430 документов в формате ‘.json’ для каждого жеста, которые содержат информацию: сколько рук обнаружено на фото и массив данных длинной 21 элемент. Массив содержит ключевые точки рук в формате x, y и z: 21 ориентир для одной руки.

2. Буквы

2.1. Буквы русского жестового языка.

Поскольку не существует такого общепризнанного набора данных, то он создан самостоятельно с помощью волонтёров. Участвовало 11 волонтёров, из которых 7 девушек и 4 мужчин в возрасте от 20 до 55 лет. Волонтёры показывали 25 букв русского алфавита (а, б, в, г, е, ж, и, к, л, м, н, о, п, р, с, т, у, ф, х, ц, ч, ы, э, ю, я) с помощью жестов. Инструкции о том, как показывать буквы жестового русского языка были взяты из проекта «Словарь. Русский жестовый язык» (https://surdo.me). Было записано 8 видео и взято одно видео из проекта «Словарь. Русский жестовый язык».

Все видео были раскадрованы и отобраны лучшие фото. В итоге, получилось 25 папок с жестами букв с n цветных фотографий на каждый жест.

Набор данных был обработан с помощью фреймворка mediapipe с открытым исходным кодом, представленный Google, который помогает создавать мультимодальные конвейеры машинного обучения, и разделён на ещё два корпуса данных.

2.1.1. 25 папок по 8 документов в формате ‘.json’ для каждого жеста, которые содержат информацию о 50 кадрах, в течение которых, показывался жест (если жест показывался меньше чем за 50 кадров, то кадры интерполировались). Массив данных длинной 50 \* 43 элемента. Массив содержит ключевые точки рук в формате x, y и z: 21 ориентир для правой руки, 21 ориентир для левой руки и точка соотношения положений двух рук относительно друг друга. Если на фотографии обнаружена одна рука, то элементы с индексами после 21 будут нулевыми.

2.1.2. 25 папок по 8 папок для каждого человека с 50 цветных фотографий для каждого жеста только обнаруженных рук размером 64 × 64.

2.2. Буквы американского жестового языка.

2.2.1. Данный набор данных найден на kaggle (<https://www.kaggle.com/datasets/kapillondhe/american-sign-language/code>) – это набор букв от a до z американского жестового языка. Из набора данных было выбрано 3000 цветных фотографий размером 400 × 400 для каждой буквы.

2.2.2. Набор данных был обработан с помощью фреймворка mediapipe с открытым исходным кодом, представленный Google, который помогает создавать мультимодальные конвейеры машинного обучения. Был создан ещё один корпус данных, который содержит 26 папок по 3000 документов в формате ‘.json’ для каждого жеста, которые содержат информацию: сколько рук обнаружено на фото и массив данных длинной 21 элемент. Массив содержит ключевые точки рук в формате x, y и z: 21 ориентир для одной руки.

**Эксперименты**

Было создано 4 модели нейронных сетей: персептрон, одномерная свёрточная нейронная сеть, сеть долгой краткосрочной памяти для ключевых точек рук и двумерная свёрточная сеть для фотографий рук.

Эксперименты проводились в три этапа. Входные данные для нейронных сетей:

* 1 этап:

- ключевые точки рук цифр русского жестового языка;

- фотографии рук цифр русского жестового языка;

- ключевые точки рук цифр американского жестового языка;

- фотографии рук цифр американского жестового языка;

* 2 этап:

- ключевые точки рук букв русского жестового языка;

- фотографии рук букв русского жестового языка;

- ключевые точки рук букв американского жестового языка;

- фотографии рук букв американского жестового языка;

* 3 этап:

- ключевые точки рук цифр и букв русского жестового языка;

- фотографии рук цифр и букв русского жестового языка;

- ключевые точки рук цифр и букв американского жестового языка;

- фотографии рук цифр и букв американского жестового языка;

Каждый набор данных разделён на три подборки – 60% - на обучение, 20% - проверочные данные и 20% - тестовая выборка.

Также для подсчёта метрик на реальном видео были сняты два видео для цифр и букв русского жестового языка и два видео для цифр и букв американского жестового языка.

**Выводы**

1. Цифры

Таблица 1 - Сравнительная таблица оценка качества модели на тестовых данных цифр

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **RSL** | | | **ASL** | | |
| precision | recall | F1-score | precision | recall | F1-score |
| Perceptron | 0.94 | 0.94 | 0.94 | 0.99 | 0.99 | 0.99 |
| CNN 1D | 0.93 | 0.92 | 0.92 | 0.99 | 0.99 | 0.99 |
| LSTM | 0.83 | 0.80 | 0.79 | 0.96 | 0.96 | 0.96 |
| CNN 2D | 0.99 | 0.99 | 0.99 | 1.00 | 1.00 | 1.00 |

Таблица 2 - Сравнительная таблица доля распознавания цифр на видео

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **RSL** | | | **ASL** | | |
| precision | recall | F1-score | precision | recall | F1-score |
| Perceptron | 0.81 | 0.79 | 0.78 | 0.39 | 0.34 | 0.30 |
| CNN 1D | 0.78 | 0.78 | 0.77 | 0.10 | 0.18 | 0.11 |
| LSTM | 0.77 | 0.72 | 0.70 | 0.14 | 0.15 | 0.12 |
| CNN 2D | 0.65 | 0.65 | 0.63 | 0.16 | 0.24 | 0.17 |

* Все модели нейронных сетей лучше обучаются на цифрах американского жестового языка;
* Модели, обучившиеся на цифрах русского жестового языка, лучше распознаются на видео в реальном времени. Скорей всего это объясняется несколькими вещами. Во-первых, что в выбранном корпусе данных не было предложено видео с показом жестов цифр для тестирования, поэтому оно снималось самостоятельно. Поскольку с американским жестовым языком я знакома плохо, то, возможно, на видео жесты показывались не совсем чётко и правильно. Во-вторых, в наборе данных ASL, на котором обучали модели, на всех фото одна и та же рука и на одном и том же чёрном фоне, то есть малое разнообразие данных. В наборе же RSL разнообразия больше, за счёт использования рук разных форм и расположение их на цветных фонах с разным уровнем освещённости;
* Лучшее качество распознавания на корпусах данных в формате ключевых точек рук;
* Модели Perceptron и CNN 1D лучше всего показали себя в распознавании цифр на видео в реальном времени.

1. Буквы

Таблица 3 - Сравнительная таблица оценка качества модели на тестовых данных букв

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **RSL** | | | **ASL** | | |
| precision | recall | F1-score | precision | recall | F1-score |
| Perceptron | 1.00 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 |
| CNN 1D | 0.98 | 0.97 | 0.97 | 1.00 | 1.00 | 1.00 |
| LSTM | 0.89 | 0.88 | 0.88 | 0.95 | 0.95 | 0.95 |
| CNN 2D | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

Таблица 4 - Сравнительная таблица доля распознавания букв на видео

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **RSL** | | | **ASL** | | |
| precision | recall | F1-score | precision | recall | F1-score |
| Perceptron | 0.45 | 0.55 | 0.46 | 0.33 | 0.41 | 0.30 |
| CNN 1D | 0.51 | 0.58 | 0.51 | 0.32 | 0.38 | 0.30 |
| LSTM | 0.09 | 0.19 | 0.09 | 0.17 | 0.21 | 0.15 |
| CNN 2D | 0.17 | 0.24 | 0.17 | 0.00 | 0.04 | 0.00 |

* На реальном видео лучше показали себя корпуса данных с ключевыми точками рук;
* Хуже всего качество модели LSTM и с помощью неё сложно распознать как жесты русских, так и английских букв на реальном видео;
* Двумерная свёрточная сеть хуже всех распознаёт буквы на реальных видео как русского, так и американского жестового языка;
* Модели Perceptron и CNN 1D лучше всего показали себя в распознавании цифр на видео в реальном времени;
* Несмотря на то, что корпус данных для американского алфавита имеет большее количество данных, модель CNN 2D очень плохо распознаёт их распознаёт на реальном видео. Возможно, дело в однообразности рук и фонов, используемых в наборе данных.

1. Цифры и буквы

Таблица 5 - Сравнительная таблица оценка качества модели на тестовых данных цифр и букв

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **RSL** | | | **ASL** | | |
| precision | recall | F1-score | precision | recall | F1-score |
| Perceptron | 0.97 | 0.97 | 0.97 | 1.00 | 1.00 | 1.00 |
| CNN 1D | 0.95 | 0.94 | 0.94 | 1.00 | 0.99 | 0.99 |
| LSTM | 0.85 | 0.85 | 0.85 | 0.88 | 0.88 | 0.88 |
| CNN 2D | 0.99 | 0.99 | 0.99 | 1.00 | 1.00 | 1.00 |

Таблица 6 - Сравнительная таблица доля распознавания цифр и букв на видео

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **RSL** | | | **ASL** | | |
| precision | recall | F1-score | precision | recall | F1-score |
| Perceptron | 0.29 | 0.36 | 0.31 | 0.22 | 0.16 | 0.15 |
| CNN 1D | 0.22 | 0.28 | 0.23 | 0.12 | 0.09 | 0.09 |
| LSTM | 0.15 | 0.14 | 0.13 | 0.08 | 0.07 | 0.06 |
| CNN 2D | 0.26 | 0.20 | 0.17 | 0.03 | 0.06 | 0.03 |

* Все модели нейронных сетей довольно плохо распознают и цифры, и буквы на реальных видео. Для видео, где показывались русские жесты, это может объяснятся тем, что количество данных, на которых обучались модели для русского жестового языка, было мало и взято мало ракурсов фотографий, а для американских жестов слишком однообразно и взято мало ракурсов;
* Все модели распознают лучше цифры и буквы русского жестового языка, чем американского. Это может объяснятся тем, что корпуса данных для цифр и букв американского жестового языка довольно однообразны в форме рук, чем наборы данных русского;
* Хуже всего показали себя модели LSTM и CNN 2D на видео.