

Распознавание слов жестового языка по видеофрагментам

Мукумов Анвар Рустамович, гр. 20122

Научный руководитель: Неделько Виктор Михайлович

Актуальность



В России, более 13 млн. человек имеют нарушения слуха, а около 300 тыс. - полностью неслышащие, и для них общение с людьми, не знающими русского жестового языка вызывает сложности и неудобства

В связи с этим, исследование в области распознавания слов языка жестов является актуальным, так как оно может помочь упростить коммуникацию людей, имеющих нарушения слуха, с окружающими, уменьшить испытываемый ими дискомфорт, в связи с их недугом

Цель исследования и значимость



Цель исследования: разработка решения задачи классификации жестов по видеофрагментам при помощи нейросетевого алгоритма, работающего на предварительно извлеченных координатах ключевых точек кистей рук в каждом кадре видеофрагментов

Исследование имеет практическую значимость, так как методы решения, представленные в имеющихся работах, опираются на данные о перемещениях других частей тела, в частности на движение лица, из-за чего модели обучаются считывать “подсказки” в эмоциях диктора и чтении слов по губам в то время как лицо диктора может частично или полностью отсутствовать в кадре. Таким образом, известные решения более чувствительны к условиям записи

Задачи

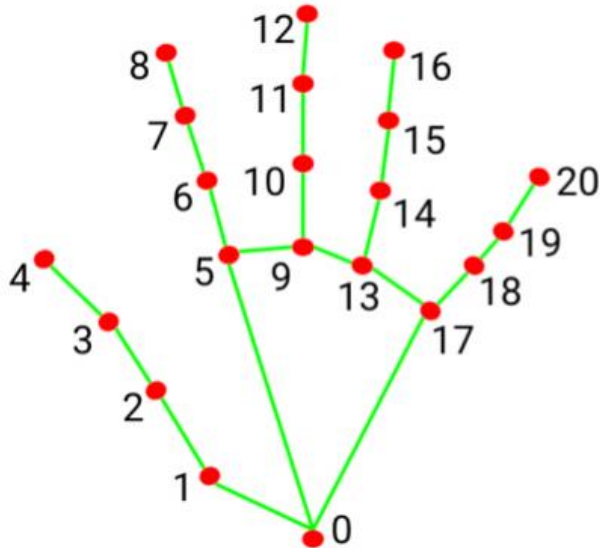


Для достижения цели были поставлены следующие задачи:

1. Построить архитектуру нейронной сети для классификации жестов по координатам ключевых точек кистей рук, извлеченных из видеофрагментов при помощи нейросети MediaPipe
2. Выбрать полезные аугментации для расширения обучающей выборки
3. Применить предобучение на стороннем наборе данных для повышения качества классификации
4. Оценить качество классификации, сравнить с имеющимися результатами

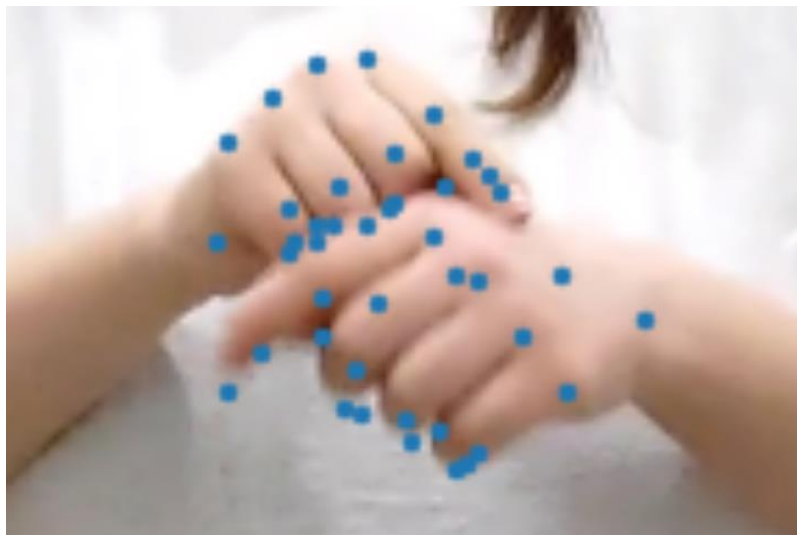
Ключевые точки

- В качестве входных данных используется 21 ключевая точка на каждой руке



0. WRIST
1. THUMB_CMC
2. THUMB_MCP
3. THUMB_IP
4. THUMB_TIP
5. INDEX_FINGER_MCP
6. INDEX_FINGER_PIP
7. INDEX_FINGER_DIP
8. INDEX_FINGER_TIP
9. MIDDLE_FINGER_MCP
10. MIDDLE_FINGER_PIP

11. MIDDLE_FINGER_DIP
12. MIDDLE_FINGER_TIP
13. RING_FINGER_MCP
14. RING_FINGER_PIP
15. RING_FINGER_DIP
16. RING_FINGER_TIP
17. PINKY_MCP
18. PINKY_PIP
19. PINKY_DIP
20. PINKY_TIP



Математическая постановка задачи



X - пространство трехмерных объектов (x_{ijk}) , $i = \overline{1, 3}$, $j = \overline{1, 42}$, $k = \overline{1, n}$ координат 42-х ключевых точек на видеофрагменте длиной в n кадров. Y - множество меток классов

Из X извлечена некоторая выборка S с известными метками классов. Требуется построить функцию $f : X \rightarrow Y$, сопоставляющую элементам из X метки класса из Y наилучшим, в некотором смысле, образом. В качестве критерия рассматривается значение выбранной метрики качества на отложенной подвыборке из S .

Обзор данных



Исследуемый набор данных - Slovo - Russian Sign Language Dataset [1, 2]. В нем имеется 20000 видеофрагментов и последовательностей координат, извлеченных покадрово с помощью нейросети MediaPipe, а также таблица с разрешениями, идентификаторами пользователей и метками классов - слов русского языка, по одному для каждого видеофрагмента. Всего имеется 2000 уникальных дикторов.

Выборка разделена на 2 части в соотношении 3:1 на обучающую и тестовую. Классы сбалансированы. Координаты x , y нормализованы к отрезку $[0, 1]$, координата z обозначает меру глубины объекта на изображении.


Степень разработанности проблемы



В открытом доступе имеются статья [1] с методами решения данной задачи, использующими видеотрегменты в качестве входных данных, с точностью классификации у лучшей модели 64%, однако в ней не была рассмотрена возможность автоматической разметки ключевых точек на руках при помощи нейронной сети, а затем обучения и формирования прогнозов на основе этих данных

Также, есть исследование [3] в области распознавания американского языка жестов, однако в нем рассмотрен гораздо больший объем данных на класс, и для обучения используются координаты ключевых точек не только кистей, но и губ, носа, и некоторых других частей тела.

Предобработка данных

- 
- Последовательности длиннее 128 кадров были обрезаны
 - Последовательности короче 128 кадров были дополнены нулями
 - На кадрах, в которых не была распознана кисть руки координаты были заменены нулем
 - Объекты преобразованы к двумерному виду $[3, 42, 128] \rightarrow [126, 128]$

Архитектура нейронной сети

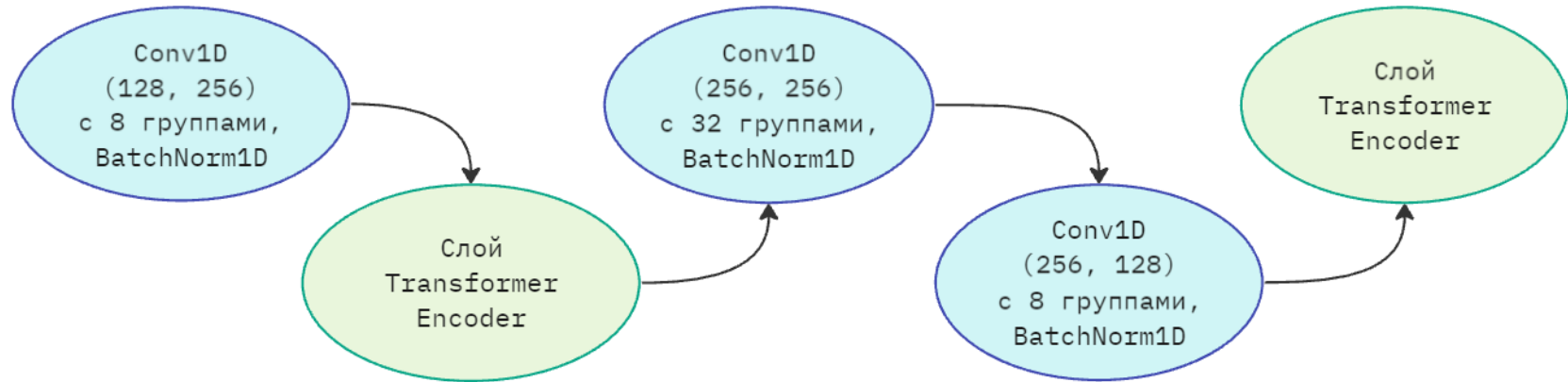
Итоговая архитектура



Архитектура нейронной сети

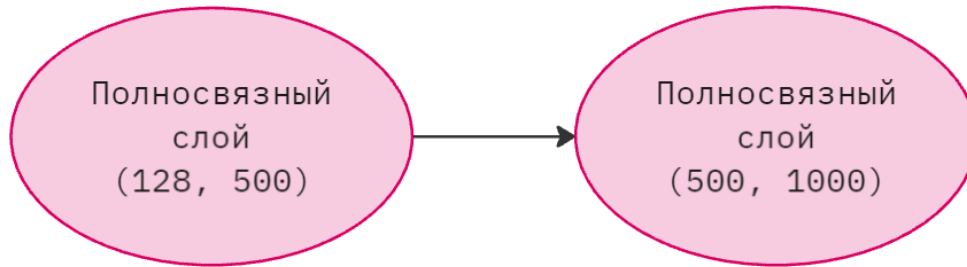


Блок генерации признаков



Архитектура нейронной сети

Классификатор



Функция потерь



В качестве функции потерь для обучения нейронной сети была выбрана перекрестная энтропия. Ее значение на объекте вычисляется по формуле:

$$H(y, p) = - \sum_{i=1}^K I\{y = t_i\} * \log(p_i)$$

где K - количество классов, t - множество меток классов, p_i - вероятность объекта принадлежать классу t_i . Затем берется среднее значение по всем объектам. В качестве оценки вероятности выступает значение функции Softmax от выходного вектора нейросети:

$$Softmax(z)_j = \frac{e^{z_j}}{\sum_{j=1}^N e^{z_j}}$$

Целевая метрика




В силу сбалансированности классов, целевой метрикой была выбрана Точность, которая вычисляется как доля верных предсказаний модели:

$$\text{Accuracy}(f, x, y) = \frac{1}{N} \sum_{i=1}^N I\{y_i = f(x_i)\},$$

где x - выборка из N объектов, y - вектор меток классов, соответствующих выборке, f - построенная решающая функция

Применяемые аугментации

- 
- Отражение относительно вертикальной оси ($p = 0.5$)
 - Сдвиг по координатным осям на случайный вектор с координатами из отрезка $[-0.5, 0.5]$ в нормализованном виде ($p = 0.35$)
 - Поворот относительно начала координат на случайный угол из отрезка $[-0.3, 0.3]$ радиан ($p = 0.35$)

Все аугментации применяются при каждом извлечении объекта перед подачей на вход в нейросеть с указанной вероятностью p

Предобучение



С целью повышения качества классификации, построенная архитектура была предобучена на наборе данных со словами американского языка жестов [4]. К данным была применена предобработка, описанная ранее, из аугментаций использовано только отражение, ввиду отсутствия информации о разрешении видеофрагментов.

На этом наборе данных была получена точность 73.8%. В качестве входных данных не использовались координаты ключевых точек губ, носа и других частей лица и тела, помимо кистей рук, несмотря на это была получена хорошая точность классификации.

Результаты



ResNet-3D	SWIN-Large	MViTv2
0.439	0.556	0.640

[1]

Модель без аугментаций	Модель с аугментациями	Предобученная модель
0.292	0.396	0.443

Результаты



В ходе проведенного исследования были получены следующие результаты:

- Построена архитектура нейронной сети, решающая задачу классификации слов русского языка жестов по координатам ключевых точек кистей рук на видеофрагментах.
- Исследовано влияние аугментаций на качество модели. Обнаружено, что в данной задаче, их применение повышает точность классификации: с 29.2% до 39.6%
- Модель предобучена на стороннем наборе данных. Выявлено, что предобучение в комбинации с аугментациями повышает точность классификации: с 39.6% до 44.3%
- Полученная точность сравнима с имеющимися результатами, однако для прогноза достаточно наличие кистей рук в кадре, в отличие от других решений данной задачи, что делает построенное решение более универсальным, устойчивым к отсутствию дополнительной информации в кадре

Источники литературы



[1] Kapitanov Alexander, Kvanchiani Karina, Nagaev Alexander, Petrova Elizaveta Slovo: Russian Sign Language Dataset. //International Conference on Computer Vision Systems (ICVS), 2023, pp. 63-76

[2] Kapitanov Alexander, Kvanchiani Karina, Nagaev Alexander, Petrova Elizaveta Slovo - Russian Sign Language Dataset // Kaggle URL: <https://www.kaggle.com/datasets/kapitanov/slovo>,

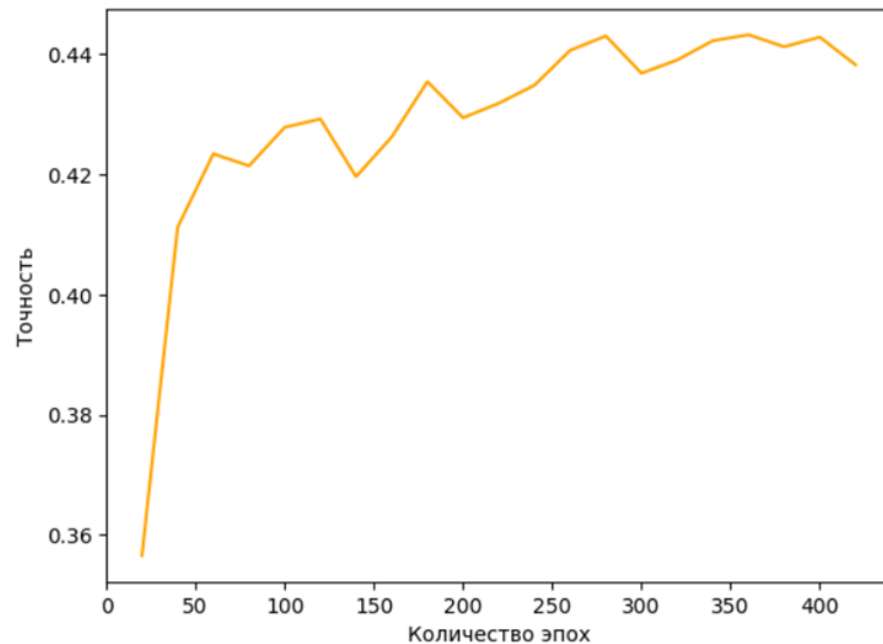
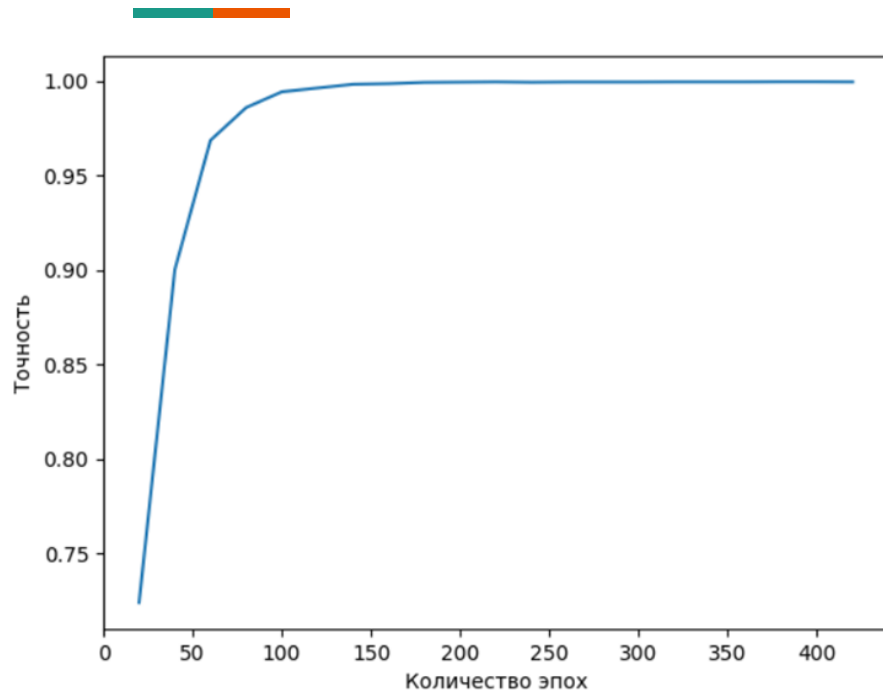
License: CC BY-SA 4.0

[3] Ahmed Mateen Buttar, Usama Ahmad, Abdu H. Gumaei, Adel Assiri, Muhammad Azeem Akbar, Bader Fahad Alkhamees Deep Learning in Sign Language Recognition: A Hybrid Approach for the Recognition of Static and Dynamic Signs // MDPI open access journal Mathematics Volume 11 Issue 17

[4] Ashley Chow, Glenn Cameron, Mark Sherwood, Phil Culliton, Sam Sepah, Sohier Dane, Thad Starner Google - Isolated Sign Language Recognition // Kaggle URL: <https://kaggle.com/competitions/asl-signs>

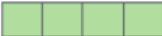
License: CC BY 4.0

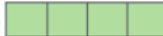
Кривая валидации




Механизм внутреннего внимания

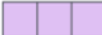
Embedding

X_1 

X_2 

Queries


q_1 

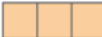
q_2 



W^Q

Keys

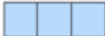
k_1 

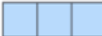
k_2 



W^K

Values

v_1 

v_2 



W^V