ПРОГРАММНОЕ СРЕДСТВО РАСПОЗНАВАНИЯ ЯЗЫКА ЖЕСТОВ

Кравченко П.Д.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Красковский П.Н. – ст.преп.

Предложена структура программного средства распознавания языка жестов для классификации American Sign Language (ASL) жестов. Описана модель, обученная на размеченных точках, полученных с помочью MediaPipe Holistic Solution.

Так как обучение модели лучше всего воспроизводить на одном языке жестов, то в качестве используемого был выбран American Sign Language (ASL) по нескольким причинам. Во-первых, он является одним из самых популярных языков жестов на данный момент времени. Во-вторых, ASL может послужить как один из тех языков, что применяются на международном уровне, чтобы носители из разных уголков планеты могли понимать друг друга. Для обучения был использован датасет, который предоставляет 250 различных жестов от 21 человека разной расы, пола, возраста, интерьера съемки [1]. Их жесты могут содержать в себе переходы с прошлых жестов или иметь какие-нибудь недочеты. Некоторые носители использовали левую руку, другие правую. Кто-то переключался с одной на другую.

Обработка входных данных будет осуществляться с помощью MediaPipe Holistic Solution [2], идея которого заключается в объединении различных моделей для рук, позы, лица человека. Использовать эти модели отдельно является не очень хорошей идеей, так как распознавание позы принимает на вход видеокадр с фиксированным разрешением (256x256), а если обрезать руки и лицо для двух других моделей, то разрешение получается слишком низким для точной артикуляции. Holistic Solution справляется с этой проблемой. Также нет необходимости использовать все 468 точек лица, поэтому остаются только те, что отвечают за область губ. Это связано с тем, что при жестикуляции человек может также проговаривать слова.

Основой программного средства является модель, которая на вход принимает данные, а на выходе возвращает значение жеста. Для решения этой задачи использованы трансформеры [3], которые были представлены в 2017 году специалистами из Google Brain с целью решения задач обработки естественного языка. Одно из основных отличий от существующих методов обработки данных заключается в том, что входная последовательность может передаваться параллельно, чтобы можно было эффективно использовать графический процессор, а также увеличивать скорость обучения.

Основными компонентами трансформеров являются энкодер и декодер. Энкодер берет на вход данные и проецирует их на пространство большей размерности (N-мерный вектор). Этот абстрактный вектор подается в декодер, который превращает его в выходную последовательность, и в данном случае обозначает значение жеста. Другие инновации, лежащие в основе трансформеров, сводятся к трем основным концепциям: позиционные энкодеры, внимание и самовнимание.

Позиционные энкодеры позволяют распараллелить процесс. Для этого используются позиционные кодировки, что помогают перенести бремя понимания порядка со структуры нейронной сети на сами данные. Сначала, прежде чем трансформеры обучатся на какой-либо информации, они не знают, как интерпретировать эти позиционные кодировки. Но по мере того, как модель видит все больше и больше примеров и их кодировок, она учится эффективно их использовать.

Механизм внимания одновременно просматривает несколько частей входной последовательности и решает, какие из них важны, приписывая им разные веса. Декодер же помимо вектора принимает и эти веса, предоставленные механизмом, что делает его работу намного проще, потому что теперь есть понимание того, что больше всего влияет на выбор выходного значения.

Последняя часть трансформеров — это поворот внимания, называемый самовниманием. Если механизм внимания определяет, какие значения вектора являются самыми важными, то самовнимание модифицирует каждое значение вектора подмешивая к нему другие близкие значения из контекста с некоторыми весами.

Список использованных источников:

1. Kaggle ASL-signs [Электронный ресурс]. – Электронные данные. – Режим доступа: https://www.kaggle.com/competitions/asl-signs/data.

2. MediaPipe Holistic Solution [Электронный ресурс]. – Электронные данные. – Режим доступа: https://github.com/google/mediapipe/blob/master/docs/solutions/holistic.md.

3. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Polosukhin I., Kaiser Ł. Attention is All you Need (англ.) // Advances in Neural Information Processing Systems 30 / I. Guyon, U. v. Luxburg, S. Bengio, H. Wallach, R. Fergus, S.V.N. Vishwanathan, R. Garnett — 2017. — arXiv:1706.03762.