АНАЛИЗ МЕТОДОВ РАСПОЗНАВАНИЯ ОБЪЕКТОВ В ВИДЕО В ЗАДАЧЕ РАСПОЗНАВАНИЯ ЖЕСТОВЫХ ЯЗЫКОВ

Д. А. Жалова

Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь, fpm.zhalova@bsu.by
Научный руководитель — В. В. Казаченок, доктор педагогических наук, профессор

Данная работа анализирует различные подходы к обработке видео с помощью глубоких нейронных сетей на примере задачи распознавания жестового языка. Рассматриваются такие архитектуры нейронных сетей, как Video Swin Transformer и последовательность из сверточных нейронных сетей и сетей долгой краткосрочной памяти. Приведены результаты сравнительного анализа моделей, выводы и рекомендации построения моделей для обработки видео.

Ключевые слова: глубокие нейронные сети; распознавание жестовых языков; компьютерное зрение; классификация видео; сверточные нейронные сети; сети долгой краткосрочной памяти.

Видеоданные имеют свои отличия от других типов данных, таких как изображения или аудио, поскольку видео представляют собой последовательность изображений, снятых с определенной частотой кадров, которые передают движение и изменения во времени. Этот тип данных имеет два типа признаков: временные признаки и пространственные и данная особенность должна учитываться при обработке видео.

Один из способов обработки видеоданных, который учитывает все их особенности – это архитектура нейронной сети Video Swin Traformer. Данная архитектура представляет собой модификацию архитектуры Swin Transformer, которая в свою очередь используется для работы с изображениями. VideoSwin основан на основных принципах трансформерных архитектур, которые изначально были успешно применены в области обработки текстовых данных.

Второй рассматриваемый метод обработки видео – это комбинация сверточных нейронных сетей и сетей долгой краткосрочной памяти. На этапе извлечения признаков свёрточной нейронной сетью используется предварительно обученная модель, в данном случае VGG-16. Далее извлеченные признаки подаются на вход слоям LSTM – сети долгой краткосрочной памяти, которая производит обработку полученных временных признаков.

Для обучения моделей нейронных сетей использовался набор данных Slovo. Данный набор содержит 20000 видеозаписей и 1000 классов, пред-

ставляющих часто используемые словосочетания и короткие фразы на русском жестовом языке, включая алфавит и цифры. В данной работе для обучения нейронных сетей использовались 20 классов жестов и дополнительный класс "no_event" со случайно выбранными 20 видео. Также к данным были применены такие методы аугментации, как отражение вдоль вертикальной оси, изменение яркости и контраста.

При реализации модели на основе VideoSwin Transformer была использована предварительно обученная модель на наборе данных Kinetics-400. На верхние слои данной модели были добавлены слой глобального среднего пулинга и полносвязные слои. В качестве функции оптимизации использовалась функция Adam со скоростью обучения 0.001. Также использовалась L2-регуляризация с параметром 0.05.

Эксперименты, проводимые с данной моделью, включали:

- изменение количества слоев на вершине модели и количества нейронов на них;
- изменение скорости обучения функции оптимизации;
- изменения размера пакета данных;
- изменение параметра регуляризации;
- изменение количества эпох.

Наилучшие результаты (рис. 1) такая архитектура показала в модификации с двумя полносвязными слоями с 32 нейронами на последнем скрытом слое. Результаты на тестовой выборке: 0.36 точность предсказания и 2.70 ошибка предсказания.

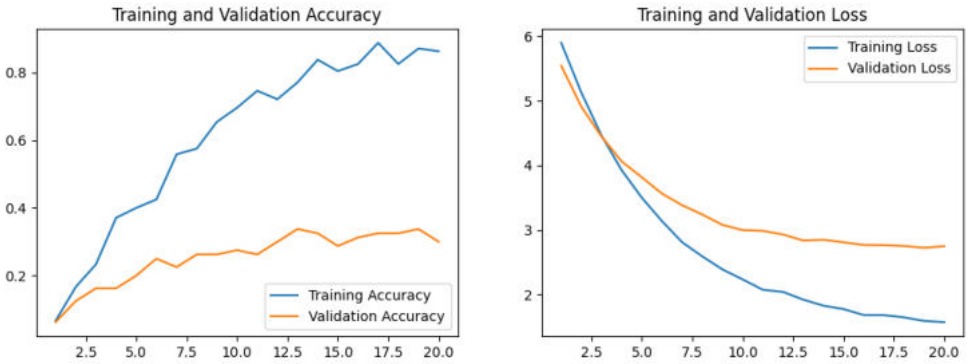

Рис. 1. Результаты модели Video Swin с двумя полносвязными слоями

Реализованная одель на основе CNN-LSTM состояла из двух главных частей: сверточной нейронной сети для обработки пространственных признаков и сети долгой краткосрочной памяти для обработки полученных признаков с учетом временных зависимостей.

Вторая часть данной модели была составлена из LSTM слоев и полносвязных слоев с разными конфигурациями. Основные вариации LSTM части данной архитектуры:

- один слой LSTM с одним полносвязным слоем после;
- один LSTM слой с двумя полносвязными слоями;
- два слоя LSTM с одним и с двумя полносвязными слоями;
- три слоя LSTM с двумя полносвязными слоями;
- два LSTM слоя.

Наилучшие результаты (рис. 2) нейронная сеть показала в случае, когда после сверточной нейронной сети следовали два LSTM слоя, со слоем Dropout между ними, а затем два полносвязных слоя. Параметр регуляризации был равен 0.05. Первые два слоя LSTM содержали 64 и 32 нейрона на выходе соответственно, а последний скрытый полносвязный слой содержал 32 нейрона на выходе. значение функции потерь на тестовой выборке составило 3.40, а значение категориальной точности – 0.10.

Рис. 2. Наилучшие результаты обучения нейронной сети на основе CNN-LSTM

На основе проведенных опытов и экспериментов с рассмотренными моделями были сделаны некоторые выводы, на которые необходимо опираться при построении эффективной модели в реальных условиях:

1. В случаях, когда для обучения нейронной сети доступно небольшое количество данных, намного лучше справляются модели на основе предварительно обученных архитектур, предназначенных для работы с видео данными, такими как VideoSwin. В случае, когда доступный набор данных имеет большое количество экземпляров данных для каких-либо определенных задач может потребоваться создать свою архитектуру для работы с видео.

2. При реализации обеих моделей были проведены эксперименты с меньшим количеством слоев, однако большим их объемом, и с большим количеством слоёв с меньшим объемом данных на каждом слое. Эксперименты с обеими архитектурами нейронных сетей показали, что

эффективнее обучаются более длинные модели с меньшим объемом данных

На основе полученных результатов можно сделать вывод, что наиболее эффективно с видеоданными работает модель на основе VideoSwin. Модель на основе CNN-LSTM также может дать хорошую точность распознавания, однако требует долгой настройки как гиперпараметров, так и различных вариаций слоев. Исходя из этого, для дальнейшего применения в реальных условиях наиболее подходит модель на основе VideoSwin, поскольку она показала наилучшие результаты. Однако для более точной и эффективной работы модели необходимо большее количество данных для обработки, что, в свою очередь, требует большего количества доступных ресурсов программного обеспечения для их обработки и хранения.

Библиографические ссылки

1. Hochreiter, S. Long Short-term Memory / S. Hochreiter, J. Schmidhuber // Neural computation. 1997. №9. P. 1735-1780.

2. Orozco, C.I. CNN-LSTM Architecture for Action Recognition in Videos / Orozco, C.I., Buemi, M.E., Berlles, J.J. 2019.

3. Slovo: Russian Sign Language Dataset / A. Kapitanov [et al.] // Computer Vision Systems: 14th International Conference, VIenna, 27-29 Sept 2023 / Springer-Verlag ; ed. Henrik I. Christensen [et al.]. Berlin, 2023. P. 63-73.

4. Video Swin Transformer / Z. Liu [et al.] // 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition / IEEE. – New Orleans, LA, USA, 2022. P. 3192-3201.