УДК 004.8

РАЗРАБОТКА НЕЙРОННОЙ СЕТИ ДЛЯ РАСПОЗНАВАНИЯ ЖЕСТОВ ДАКТИЛЬНОЙ АЗБУКИ

Кобзарь Д. С. (ИТМО)
Научный руководитель – кандидат технических наук, доцент ФПИиКТ Кугаевских А. В. (ИТМО)

Введение. Жестовый язык представляет из себя совокупность жестов и мимики, соответствующих определенным словам и понятиям, и является основной формой коммуникации для людей с нарушениями слуха. Вместе с тем процент владения жестовыми языками среди населения слишком мал, что создает коммуникационные барьеры для слабослышащих людей во всех сферах жизни. В связи с тем, что все жестовые языки уникальны и имеют свои особенности, использование одной и той же модели для разных языков является неэффективным. В настоящее время существует не так много систем для распознавания жестовых языков в целом, соответственно для русского языка их еще меньше. Основная часть предложенных решений либо спроектированы с учетом использования различных сенсоров, что повышает их себестоимость и сложность, либо – используют большие модели, требовательные к ресурсам [1, 2]. Таким образом, построение более простой и быстрой модели для распознавания русского жестового языка имеет большое практическое значение. Актуальность эффективного и небольшого, а следовательно, и более доступного решения, обуславливается как социальной значимостью, так и заинтересованностью бизнеса в расширении целевой аудитории среди людей с нарушениями слуха.

Основная часть. В качестве архитектуры модели для распознавания жестов было решено рассмотреть следующие:

1) YOLO;
2) ViT (Vision Transformer) - YOLOS.

YOLO имеет сверточную архитектуру и полносвязные слои на выходе и является популярной моделью для решения задачи детектирования, обеспечивая высокую скорость работы и точность. Также к преимуществам YOLO стоит отнести простоту модели и нетребовательность к ресурсам для работы, а также хорошие способности обобщения, что позволяет упростить дообучение [3].

Модель ViT появилась вследствие успеха архитектуры трансформера для работы с текстовыми последовательностями, благодаря возможности хорошо находить зависимости между частями предложений, расположенным далеко друг от друга. Оказалось, что если разбить изображение на части и рассматривать их как последовательность токенов, то можно получить хорошие результаты и для обработки изображений. Для того, чтобы перейти от задачи классификации к задаче детектирования авторы модели YOLOS внесли всего несколько изменений: убрали классификационный токен и добавили 100 токенов для детекции, в остальном модель максимально повторяет архитектуру визуального трансформера [4]. Преимуществом по сравнению с YOLO в данном случае является как раз способность находить глобальные закономерности на изображении.

Выводы. Рассмотренные модели были реализованы и обучены на имеющемся наборе данных для русского жестового языка, по результатам экспериментов все модели показали высокое качество детектирования по следующим метрикам: precision, recall, mAP50 и mAP50-95. К наиболее точным в результате можно отнести: YOLOS (ViT) и YOLOv3-tiny-u; по времени инференса модель YOLOv3-tiny-u оказалась существенно эффективнее.

Список использованных источников:

1. Kumar R. et al. A comparative analysis of techniques and algorithms for recognising sign language //arXiv preprint arXiv:2305.13941. – 2023.

2. Shi B. Toward American Sign Language Processing in the Real World: Data, Tasks, and Methods //arXiv preprint arXiv:2308.12419. – 2023.

3. Redmon J. et al. You only look once: Unified, real-time object detection //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2016. – С. 779-788.

4. Fang Y. et al. You only look at one sequence: Rethinking transformer in vision through object detection//Advances in Neural Information Processing Systems. – 2021. – Т. 34. – С. 26183-26197.