УДК 004.932.2

«МАТЕМАТИЧЕСКИЕ ОСНОВЫ АРХИТЕКТУРЫ YOLOV8 ДЛЯ ДЕТЕКЦИИ ОБЪЕКТОВ»

Агитаев Исамурад Амандыкулы
colesattac@gmail.com
магистрант 2 курса образовательной программы «Информационные системы»
Торайгыров университет, г. Павлодар, Республика Казахстан
Научный руководитель – кандидат педагогических наук, доцент Найманова Д.С.

Задача детекции объектов на изображениях является одной из центральных проблем математической кибернетики и теории распознавания образов. На протяжении последних десятилетий данная область демонстрирует стремительное развитие: от классических методов, основанных на ручном конструировании признаков, до современных глубоких нейронных сетей, способных автоматически извлекать информативные представления из необработанных пикселей. Ранние подходы к детекции – метод скользящего окна в сочетании с дескрипторами HOG (Histogram of Oriented Gradients) и классификаторами SVM – требовали значительных вычислительных ресурсов и обладали ограниченной способностью к обобщению [3]. Практическая значимость задачи определяется широтой областей применения: системы автономного вождения, медицинская диагностика, промышленный контроль качества, системы безопасности и ассистивные технологии для людей с нарушениями зрения [1]. Переломным моментом в истории задачи детекции стало появление двухэтапных детекторов семейства R-CNN, в которых глубокие свёрточные сети впервые были применены для извлечения признаков из кандидатов регионов. Дальнейшее развитие привело к появлению одноэтапных детекторов, объединивших предсказание регионов и классификацию в единый проход через сеть. Среди одноэтапных детекторов семейство YOLO (You Only Look Once) занимает особое место благодаря уникальному сочетанию высокой скорости инференса и конкурентоспособной точности. Архитектура YOLOv8, представленная компанией Ultralytics в 2023 году, является актуальным представителем данного семейства и реализует ряд математически обоснованных улучшений по сравнению с предшественниками [2]. В данной работе рассматриваются математические основы архитектуры YOLOv8 в контексте разработки системы распознавания объектов для помощи слабовидящим людям.

Математическая природа свёрточной операцииФундаментальной вычислительной единицей в архитектуре YOLOv8 является дискретная двумерная свёртка – операция, имеющая строгое математическое определение в теории сигналов и функционального анализа. Входное изображение представляется в виде трёхмерного тензора, где два измерения соответствуют пространственным координатам, а третье – цветовым каналам. Применение набора обучаемых фильтр-ядер к этому тензору порождает карты признаков, каждая из которых фиксирует реакцию соответствующего фильтра на локальные паттерны в разных областях изображения [1, с. 326-330]. Ключевое свойство свёрточной операции, которое обеспечивает её эффективность для обработки изображений – это разделение параметров. То есть, один и тот же фильтр будет применяться ко всем пространственным позициям входного тензора. Таким образом обеспечивается трансляционная инвариантность – способность обнаружить один и тот же признак, независимо от его расположения в изображении. Также это значительно сокращает число параметров обучения, по сравнению с полносвязными слоями. Если полносвязный слой для изображения 640x640x3 потребует свыше миллиона параметров только для одного нейрона выходного слоя, то свёрточный фильтр размером 3x3x3 содержит всего 27 обучаемых параметров, вне зависимости от размеров входного изображения [2].

Иерархия представлений формируется посредством последовательного применения нескольких свёрточных слоёв. Нижние слои извлекают элементарные геометрические примитивы, такие как рёбра, углы, градиенты яркости; средние слои формируют уже более сложные паттерны: текстуры, части объектов; верхние слои генерируют семантически значимые признаки, это формы объектов и их взаимное расположение. Именно эта иерархическая природа свёрточных сетей обусловливает их высокую эффективность в задачах компьютерного зрения [1, с. 326–330]. Экспериментально подтверждено, что признаки нижних слоёв глубоких сетей универсальны и хорошо переносятся между задачами, тогда как признаки верхних слоёв специфичны для конкретной задачи и набора данных – это явление лежит в основе метода переноса обучения (transfer learning), широко применяемого в задачах детекции [4]. После свёрточных слоёв применяются нелинейные функции активации. В современных архитектурах наиболее распространена функция ReLU и её модификации (Leaky ReLU, SiLU), обеспечивающие вычислительную эффективность и способность сети аппроксимировать нелинейные зависимости произвольной сложности [1, с. 168–171]. Функция SiLU, применяемая в YOLOv8, представляет собой произведение аргумента и сигмоиды от аргумента, что обеспечивает плавную нелинейность и более стабильные градиенты по сравнению с ReLU, особенно в области малых отрицательных значений. Пулинговые операции выполняют субдискретизацию карт признаков, снижая их пространственное разрешение и придавая модели инвариантность к малым сдвигам и деформациям объектов [3].

Трёхкомпонентная архитектура YOLOv8 организована в виде трёх функционально различных компонентов: backbone, neck и head, каждый из которых выполняет строго определённую математическую роль в процессе детекции [2]. Backbone на основе CSPDarknet (Cross Stage Partial Network) отвечает за извлечение многомасштабного представления входного изображения. Идея CSP состоит в разделении входного тензора каждого блока на две части: одна проходит через последовательность свёрточных слоёв, другая напрямую передаётся к выходу блока, после чего обе части объединяются конкатенацией. Это сокращает вычислительную сложность при сохранении репрезентативной способности сети. Ключевым математическим элементом backbone являются остаточные связи (residual connections), концепция которых была введена в архитектуре ResNet [4]. Суть остаточного блока состоит в том, что к выходу последовательности нелинейных преобразований добавляется необработанный вход блока. С точки зрения теории оптимизации это обеспечивает беспрепятственное прохождение градиентов при обратном распространении ошибки через сотни слоёв, решая проблему затухания градиентов [4]. Neck реализует механизм объединения признаков с разных уровней иерархии посредством Feature Pyramid Network (FPN) и Path Aggregation Network (PAN).

FPN строит пирамиду признаков внушительного масштаба, распространяя семантически богатую информацию с верхних уровней на нижние путём поэлементного сложения тензоров после их билинейной интерполяции до нужного пространственного разрешения. Восходящим путём агрегации PAN дополняет этот процесс и обеспечивает точную локализационную информацию с нижних уровней для верхних [5]. Совместное применение этих механизмов позволяет детектировать объекты одновременно на нескольких пространственных масштабах: крупные объекты детектируются на картах признаков малого пространственного разрешения с большим рецептивным полем, мелкие – на картах высокого разрешения. Это принципиально важно для ассистивных систем, где размеры и расстояния до объектов варьируются в широком диапазоне. Head выполняет итоговые предсказания. В anchor-based детекторах предсказания формируются относительно набора предопределённых якорных рамок, что требует тщательного подбора их параметров применительно к конкретному набору данных и усложняет математическую постановку задачи. Также, anchor-based подходы вносят дополнительную

неопределённость при сопоставлении предсказаний с истинными разметками в процессе обучения.

YOLOv8 напрямую регрессирует координаты ограничивающего прямоугольника для каждой точки выходной сетки, устраняя зависимость от якорных рамок и упрощая алгоритм сопоставления предсказаний с истинными разметками [2]. Для задания соответствия между предсказаниями и истинными боксами в процессе обучения применяется алгоритм Task-Aligned Assigner, совместно учитывающий качество классификации и точность локализации.

Обучение YOLOv8 формализуется как задача минимизации составной функции потерь, включающей три аддитивных компонента с весовыми коэффициентами: потери классификации, потери регрессии координат боксов и Distribution Focal Loss [2]. Весовые коэффициенты являются гиперпараметрами, подобранными экспериментально на стандартных тестовых наборах данных. Суммарная функция потерь суммируется по всем точкам выходной сетки и по всем масштабам, на которых выполняется детекция. Потери классификации основаны на бинарной кросс-энтропии – фундаментальной функции потерь теории информации. Она измеряет расхождение между истинным распределением меток классов и предсказанным распределением вероятностей. С информационно-теоретической точки зрения бинарная кросс-энтропия представляет собой оценку числа бит, необходимых для кодирования истинных меток с использованием предсказанного распределения вместо истинного [1, с. 178–183]. Применяя бинарную кросс-энтропию вместо категориальной, можно рассматривать задачу определения класса объекта как набор независимых бинарных классификационных подзадач. Это будет математически эквивалентно предположению о независимости меток классов. Потери регрессии боксов основаны на метрике CIoU (Complete Intersection over Union). Исходная метрика IoU – отношение площади пересечения к площади объединения двух прямоугольников – является геометрически интерпретируемой мерой качества локализации, инвариантной к масштабу объекта. Но стандартный IoU имеет и свои недостатки как функция потерь. Он равен нулю при отсутствии пересечения боксов, из-за этого даже при значительном расхождении предсказания от истины, он приводит к нулевому градиенту. GIoU (Generalized IoU) частично устраняет этот недостаток. Он вводит штраф на основе наименьшего охватывающего прямоугольника. CIoU развивает эту идею дальше, дополнительно учитывая евклидово расстояние между центрами боксов и штраф за несоответствие их соотношений сторон [6]. Благодаря этому мы получаем более стабильную и информативную оптимизацию, особенно для объектов малого размера, критически важных в реальных сценариях использования.

Distribution Focal Loss (DFL) – это математически нетривиальный подход к задаче регрессии координат. Она не предсказывает точечные значения координат, вместо этого модель предсказывает дискретное вероятностное распределение над возможными значениями каждой координаты, из которого затем извлекается ожидаемое значение. С таким подходом можно явно моделировать неопределённость в предсказании координат, что особенно важно для размытых или частично перекрытых объектов [7]. Применение Focal Loss позволяет сосредоточить обучение на трудных примерах через уменьшение вклада легко классифицируемых фоновых точек в суммарную функцию потерь. Качество детекции оценивается через метрику mean Average Precision (mAP), которая вычисляется как среднее значение площади под кривой Precision-Recall по всем классам объектов при заданном пороге IoU. Значение mAP@50 соответствует порогу IoU=0,5, а mAP=50-95 усредняет результаты по диапазону порогов от 0,5 до 0,95 с шагов 0,05 и обеспечивает более строгую и объективную оценку качества локализации [3].

Стохастическая оптимизация и регуляризация. Минимизация функции потерь осуществляется методами стохастической оптимизации первого порядка. В YOLOv8 применяются SGD с импульсом и адаптивный метод Adam. SGD с импульсом накапливает экспоненциально взвешенное среднее градиентов, благодаря чему возможно двигаться по

пологим направлениям потерь с ускорением и демпфировать осцилляции вдоль крутых направлений. Adam расширяет этот принцип, дополнительно адаптируя скорость обучения индивидуально для каждого параметра на основе оценок первого и второго статистических моментов градиента [1, с 306-310]. Коррекция смещения первых итераций, введённая в алгоритме Adam, обеспечивает корректные оценки моментов на начальных шагах обучения, когда накопленная история градиентов мала. На практике же, для обучения YOLOv8 рекомендуется использовать SGD с косинусным расписанием скорости обучения, постепенно снижающим шаг оптимизации по косинусному закону от начального значения до нуля на протяжении всего обучения. Для предотвращения переобучения в YOLOv8 применяется комплекс мер. Регуляризация весов (weight decay) соответствует L2-регуляризации в пространстве параметров и штрафует большие значения весов, стимулируя модель к нахождению более простых решений. Mosaic-аугментация, введённая в YOLOv4 и сохранённая в YOLOv8, стохастически объединяет четыре обучающих изображения в одно, существенно расширяя эффективный объём обучающей выборки и вынуждая модель детектировать объекты в непривычном контексте и при нестандартном масштабе [2]. MixUp-аугментация интерполирует пары изображений и их разметок с случайным коэффициентом, что формирует более мягкие границы принятия решений. Также применяются случайные геометрические преобразования (горизонтальные и вертикальные отражения, масштабирование, сдвиги, перспективные искажения) и фотометрические искажения (изменения яркости, контраста, насыщенности, оттенка). Совокупность этих методов регуляризации позволяет модели обобщаться на неизвестные данные значительно лучше, чем при обучении без аугментации.

Применение YOLOv8 в ассистивных технологиях. В разрабатываемой системе Vision Assist архитектура YOLOv8 выступает математической основой модуля детекции объектов реального мира. Система реализована в виде Progressive Web Application (PWA) с Python-бэкендом на базе FastAPI, что обеспечивает кроссплатформенную доступность без необходимости установки специализированного программного обеспечения. Входное изображение с камеры мобильного устройства стандартизируется до размера 640×640 пикселей с сохранением пропорций путём добавления серых полос (letterboxing) и нормализацией значений пикселей к диапазону [0, 1].Нормализованное изображение подаётся в backbone, формирующий многомасштабный тензор признаков на трёх пространственных масштабах: 80×80, 40×40 и 20×20 ячеек, соответствующих мелким, средним и крупным объектам соответственно. По результатам инференса система генерирует набор обнаруженных объектов: для каждого объекта определяются координаты ограничивающего прямоугольника, класс из 80 предопределённых категорий датасета COCO и значение достоверности (confidence score). Объекты с достоверностью ниже заданного порога отфильтровываются, после чего применяется алгоритм Non-Maximum Suppression (NMS) для устранения дублирующих предсказаний путём последовательного отбора предсказаний с наивысшей достоверностью и подавления всех перекрывающихся с ними предсказаний выше порога IoU [8]. Итоговый список распознанных объектов передаётся в модуль синтеза речи на основе Web Speech API для голосового оповещения пользователя. Система поддерживает три языка: русский, казахский и английский, что обеспечивает её актуальность для полиязычной аудитории Казахстана и отличает её от большинства существующих коммерческих решений, ориентированных исключительно на английский язык. Дополнительно система включает модуль распознавания текста на основе Tesseract OCR, позволяющий зачитывать вслух надписи, вывески и этикетки, обнаруженные на изображении [9].Принципиальным преимуществом YOLOv8 для данного приложения является возможность использования облегчённой версии YOLOv8n. Данная версия содержит значительно меньше параметров по сравнению с полноразмерной моделью и обеспечивает приемлемую скорость инференса на мобильных процессорах без аппаратного ускорения, достигая частоты обработки кадров, достаточной для работы в

режиме реального времени. Это делает систему доступной для широкого круга конечных пользователей без необходимости в специализированном оборудовании [10].

Заключение

Рассмотренный математический аппарат архитектуры YOLOv8 включает несколько взаимосвязанных компонентов. Иерархическое извлечение пространственных признаков посредством дискретной свёртки с разделением параметров формирует основу для понимания визуальных сцен. Трёхкомпонентная структура backbone–neck–head с остаточными связями, механизмом CSP и многомасштабной агрегацией признаков посредством FPN и PAN обеспечивает эффективную детекцию объектов различных размеров. Составная функция потерь, объединяющая бинарную кросс-энтропию, CIoU-регрессию и DFL, формулирует задачу обучения как математически корректную задачу оптимизации с информативными градиентами. Адаптивные методы стохастической оптимизации в сочетании с комплексом методов регуляризации и аугментации данных обеспечивают сходимость модели к обобщающим решениям.

Список использованной литературы:

1. Goodfellow I., Bengio Y., Courville A. Deep learning. — Cambridge: MIT Press, 2016. — 775 с.
2. Jocher G., Chaurasia A., Qiu J. Ultralytics YOLOv8. — GitHub, 2023. — URL: https://github.com/ultralytics/ultralytics (дата обращения: 01.03.2025).
3. Everingham M., Van Gool L., Williams C. K. I., Winn J., Zisserman A. The Pascal Visual Object Classes (VOC) Challenge — International Journal of Computer Vision. — 2010. — Vol. 88, № 2. — С. 303-338.
4. He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition // Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). — 2016. — С. 770-778.
5. Lin T.-Y., Dollár P., Girshick R., He K., Hariharan B., Belongie S. Feature pyramid networks for object detection // Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). — 2017. — С. 2117-2125.
6. Zheng Z., Wang P., Liu W., Li J., Ye R., Ren D. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression // Proceedings of the AAAI Conference on Artificial Intelligence. — 2020. — Vol. 34, № 07. — С. 12993-13000.
7. Li X., Wang W., Wu L., Chen S., Hu X., Li J., Tang J., Yang J. Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection // Advances in Neural Information Processing Systems (NeurIPS). — 2020. Vol. 33. — С. 21002-21012.
8. Redmon J., Divvala S., Girshick R., Farhadi A. You only look once: Unified, real-time object detection // Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). — 2016. — С. 779-788.
9. LeCun Y., Bengio Y., Hinton G. Deep learning // Nature. — 2015. Vol. 521, № 7553. — С. 436-444.
10. Tan M., Le Q. V. EfficientNet: Rethinking model scaling for convolutional neural networks // International conference on machine learning (ICML). — 2019. — С. 6105-6114.