УДК 004.89
https://doi.org/10.37661/1816-0301-2022-19-3-74-85

Оригинальная статья
Original Paper

Распознавание изображений товаров электронной коммерции с использованием модели внимания и нейронной сети YOLACT

В. В. Сорокина[1]✉, С. В. Абламейко[1, 2]

[1]*Белорусский государственный университет,*
пр. Независимости, 4, Минск, 220050, Беларусь
✉*E-mail: viktoria.sorokina.96@gmail.com*

[2]*Объединенный институт проблем информатики*
Национальной академии наук Беларуси,
ул. Сурганова, 6, Минск, 220012, Беларусь

Аннотация

Цели. Предлагается алгоритм распознавания изображений товаров электронной коммерции с использованием модели внимания и нейронной сети YOLACT. Целью работы является улучшение взаимодействия между перекрестными признаками изображения с помощью модульной архитектуры, в которой применяется модель внимания к разным веткам сети.

Методы. Основными методами распознавания изображений товаров электронной коммерции являются создание и аннотация набора данных для обучения нейронной сети, выбор архитектуры и встраивание модели внимания, валидация и проведение тестов, а также интерпретация результатов.

Результаты. Сверточная нейронная сеть YOLACT модифицировалась моделью внимания для решения задачи распознавания объектов электронной коммерции, что позволило получить более качественные результаты, чем у классической сети YOLACT.

Заключение. В ходе эксперимента был подготовлен набор данных товаров электронной коммерции, произведена его аннотация, построены две нейронные сети для сравнения результатов. Результаты исследования показали, что использование модели внимания положительно влияет как на качество обученной сети, так и на скорость сходимости. Это отражается в улучшенных метриках для распознавания и сегментации объектов.

Ключевые слова: распознавание объектов, сверточная нейронная сеть, модель внимания, сеть YOLACT, электронная коммерция

Для цитирования. Сорокина, В. В. Распознавание изображений товаров электронной коммерции с использованием модели внимания и нейронной сети YOLACT / В. В. Сорокина, С. В. Абламейко // Информатика. – 2022. − Т. 19, № 3. – С. 74–85. https://doi.org/10.37661/1816-0301-2022-19-3-74-85

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию | Received 12.06.2022
Подписана в печать | Accepted 18.08.2022
Опубликована | Published 29.09.2022

E-commerce image recognition using attention model and YOLACT neural network

Viktoria V. Sorokina[1]✉, Sergey V. Ablameyko[1, 2]

[1]Belarusian State University,
av. Nezavisimosti, 4, Minsk, 220050, Belarus
✉E-mail: viktoria.sorokina.96@gmail.com

[2]The United Institute of Informatics Problems
of the National Academy of Sciences of Belarus,
st. Surganova, 6, Minsk, 220012, Belarus

Abstract

Objectives. We propose the algorithm for e-commerce image recognition using attention model and neural network YOLACT. A modular architecture is used that applies an attention model to different branches of the network in order to improve the interaction between image cross-features.

Methods. The main methods to recognize e-commerce products are the creation and annotation of a dataset for the neural network training, the choice of architecture and embedding an attention model, the validation and testing, and interpretation of the results.

Results. Convolutional neural network YOLACT has been modified by the attention model to solve image recognition task that allowed to obtain results superior in quality to the results showed by classic YOLACT.

Conclusion. In the course of the experiment, a data set of e-commerce products was prepared, annotated, and two neural networks were built to compare the results. The results of the study showed that the use of the attention model has a positive effect on both the quality of the trained network and on the rate of convergence, which is reflected in improved metrics for object recognition and segmentation.

Keywords: object recognition, convolutional neural network, attention model, network YOLACT, e-commerce

For citation. Sorokina V. V., Ablameyko S. V. *E-commerce image recognition using attention model and YOLACT neural network*. Informatika [*Informatics*], 2022, vol. 19, no. 3, pp. 74–85 (In Russ.).
https://doi.org/10.37661/1816-0301-2022-19-3-74-85

Conflict of interest. The authors declare of no conflict of interest.

Введение. На сегодняшний день распознавание объектов является одной из ключевых задач компьютерного зрения. За последние несколько лет появилось достаточное количество различных подходов к решению данной задачи. Распознавание объектов предполагает определение локализации и класса объекта на изображении. Алгоритмы создают список категорий объектов, присутствующих на изображении вместе с выровненной по осям ограничивающей рамкой (bounding box), указывающей положение и масштаб каждого экземпляра каждой категории объектов. Распознавание объектов играет важную роль в широком спектре приложений, включая анализ медицинских изображений, автономные транспортные средства, видеонаблюдение и дополненную реальность, а также в сфере электронной коммерции.

В виртуальном мире электронной коммерции фотография товара имеет ключевое значение. Продажи интернет-магазинов в значительной степени зависят от внешнего вида товаров. Важно использовать качественные изображения для электронной коммерции, чтобы привлечь трафик, визуально ответить на вопросы и превратить посетителей сайта в покупателей.

Основными требованиями к фотографиям товаров электронной коммерции на сегодняшний день являются следующие:

– размер минимум 500×500 пикселов. Так, например, требования сети Amazon – 1000×1000, а Walmart – 2000×2000 пикселов;

– формат TIFF, JPEG, PNG, JPG (наиболее популярный);

– соотношение сторон 1:1, однако для определенных категорий товаров может понадобиться портретная ориентация;

– высококачественные снимки с фокусом на продукте и профессиональным освещением;

– большинство продавцов предпочитают белый фон;
– площадь, занимаемая продуктом в кадре, как минимум 50 %;
– разрешение 72–300 dpi.

Процесс подготовки изображения товара электронной коммерции является трудозатратным с точки зрения как материальных, так и человеческих ресурсов. Для экономии многие продавцы, особенно представители малого и среднего бизнеса, снимают продукты без привлечения профессиональных фотографов и аренды студий, а затем обрабатывают фотографии в различных редакторах. Таким образом, необработанные фотографии товаров электронной коммерции могут отличаться по качеству, хотя и имеют общую специфику.

Существует множество алгоритмов для распознавания объектов. В более традиционных подходах используются алгоритмы компьютерного зрения для определения различных характеристик изображения, таких как цветовая гистограмма или края, и для идентификации групп пикселов, которые могут принадлежать объекту. Эти результаты затем передаются в регрессионную модель, предсказывающую местоположение объекта вместе с его меткой.

Вместе с тем подходы, основанные на глубоком обучении, используют сверточные нейронные сети (англ. convolutional neural network, CNN), в которых признаки объекта не нужно определять и извлекать отдельно. Поскольку методы глубокого обучения являются передовыми для задачи распознавания объектов, то именно на них будет сосредоточено внимание в данной работе.

Как уже упоминалось выше, распознавание объектов можно сформулировать как задачу нахождения объекта на изображении, т. е. его локализации, выраженной в определении рамки (bounding box), и классификации – присвоения метки для определенного объекта.

В статье предлагается подход для распознавания изображений товаров электронной коммерции, работающий в реальном времени и основанный на использовании архитектуры YOLACT [1], которая модифицирована при помощи модели внимания [2].

Анализ существующих подходов. Важность изображений в электронной коммерции хорошо изучена. В настоящее время применение нейронных сетей для распознавания товаров электронной коммерции в основном охватывает следующие две задачи:

– классификация изображений. Это фундаментальная задача компьютерного зрения, которая стремится разделить изображения на разные категории;

– распознавание объекта, т. е. определение объекта на изображении путем указания обрамляющей прямоугольной рамки (bounding box) и его дальнейшая категоризация. За последние несколько лет в связи с продолжающимся развитием нейронных сетей многие ученые и разработчики создали и оптимизировали такие фреймворки, как Caffe, TensorFlow, MXNet и PyTorch, чтобы помочь ускорить процедуры обучения и прогнозирования.

В данной работе рассматривается распознавание изображений товаров электронной коммерции как конкретный исследовательский вопрос, связанный с задачей распознавания объектов. В настоящее время методы компьютерного зрения уже получили широкое распространение при решении задач распознавания объектов, однако для распознавания изображений товаров электронной коммерции применяются гораздо реже. Задача распознавания изображений товаров электронной коммерции является более сложной, чем обычное обнаружение объектов. Несмотря на то что она требует учета некоторых специфических ситуаций, при ее решении используются те же методы.

В работе [3], посвященной построению умной системы для подбора оптимальных изображений товаров в электронной коммерции, для классификации изображений используются два типа алгоритмов: на основе методов машинного обучения, например случайный лес, и неглубокие сети в сочетании с сетями типа ResNet в качестве основы. Недостатком данного подхода является высокая чувствительность модели к обучающему множеству.

В исследовании [4] для классификации изображений электронной коммерции на основе контента был предложен алгоритм, использующий метод кластеризации *k*-средних с вычислением расстояния между классами изображений. Построенная модель показала высокую точность классификации, однако требует предварительной обработки изображений в виде удаления шумов, выравнивания цветов и сегментации самих объектов.

Лукас Боссард и др. [5] предложил классификацию одежды и разработал для этого набор данных из более чем 80 000 изображений, применяя алгоритмы случайного леса, трансдуктивный метод опорных векторов (англ. Transductive Support Vector Machine, TSVM) и трансферный лес.

В работе [6] разработана улучшенная модель распознавания с использованием региональных сверточных нейронных сетей (R-CNN). Модель обучали поверх сверточной сети AlexNet и использовали веса предварительно обученной сети ImageNet. Недостатками модели являются сложность реализации, невозможность достичь скорости реального времени и применение нестандартных слоев.

В настоящей статье используется сеть YOLACT, которая относится к классу CNN. В архитектуре CNN веса являются элементами ядра – матрицы, участвующей в операции свертки. Каждое из ядер «скользит» по соответствующим входным каналам изображения, создавая обработанную их версию. Отличительной чертой YOLACT является скорость, на момент представления это был самый быстрый метод распознавания объектов и сегментации экземпляров в реальном времени [1]. Однако в связи с тем что YOLACT является полностью сверточной нейронной сетью, ее обучение происходит медленно и сложно, поскольку каждый отдельный этап нужно обучать отдельно. Для решения этой проблемы авторы предлагают использовать модель внимания.

Основная цель работы заключается в повышении предсказательной точности модели с сохранением скорости реального времени. Для этого решается задача распознавания изображений товаров электронной коммерции, необходимо распознать на изображениях объекты 26 классов.

Выбор нейронной сети. В зависимости от архитектуры и подхода к распознаванию объектов можно выделить следующие виды нейронных сетей:

– полностью сверточные сети (fully convolutional networks, FCN);

– сверточные сети с графическими моделями (convolutional models with graphical models);

– модели на основе кодера-декодера (encoder-decoder based models);

– модели на основе многомасштабных и пирамидальных сетей (multi-scale and pyramid network based models);

– региональные сверточные сети (R-CNN based models);

– расширенные сверточные модели и семейство DeepLab (dilated convolutional models and DeepLab family);

– рекуррентные сверточные сети (recurrent neural network based models);

– генеративные модели и состязательное обучение (generative models and adversarial training);

– сверточные модели с активными контурами (convolutional models with active contour models).

Все они имеют преимущества и недостатки в зависимости от исходных требований. На выбор архитектуры в данном исследовании влияли два фактора: скорость предсказания (работа должна осуществляться в режиме реального времени) и способность к поддержке распознавания внутриклассовой вариации. Объекты из подобных подчиненных категорий часто имеют лишь незначительные различия.

В связи с тем что нейронная сеть YOLACT являлась самой быстрой моделью для распознавания объектов, была выбрана именно ее архитектура. Эта модель позволяет решать задачи распознавания объектов и сегментации с полностью сверточной топологией. В представленной работе она используется для распознавания изображений товаров электронной коммерции в реальном времени.

Основная идея авторов при создании классической архитектуры YOLACT [1] заключается в добавлении ветви маски к существующей одноэтапной модели для решения задачи сегментации аналогично добавлению ветви маски в нейронную сеть Faster R-CNN при создании нейронной сети Mask R-CNN, но без явного шага локализации функции (например, повторного объединения функций). Для этого задача сегментации была разделена на две более простые параллельные задачи, результаты которых могут быть объединены для формирования финальных масок. Первая ветвь использует FCN для создания набора масок прототипа размером, совпадающим с самим изображением, которые не зависят от каких-либо экземпляров распознаваемых

объектов. Вторая ветвь добавляется к ветви обнаружения объекта, чтобы предсказать вектор коэффициентов маски для каждого якоря, кодирующего представление экземпляра в пространстве прототипа. Наконец, для каждого экземпляра, который остается после прохождения через алгоритм NMS (non maximum suppression, техника максимального подавления), создается маска данного экземпляра путем линейной комбинации этих двух ветвей. Архитектура YOLACT показана на рис. 1 [1].

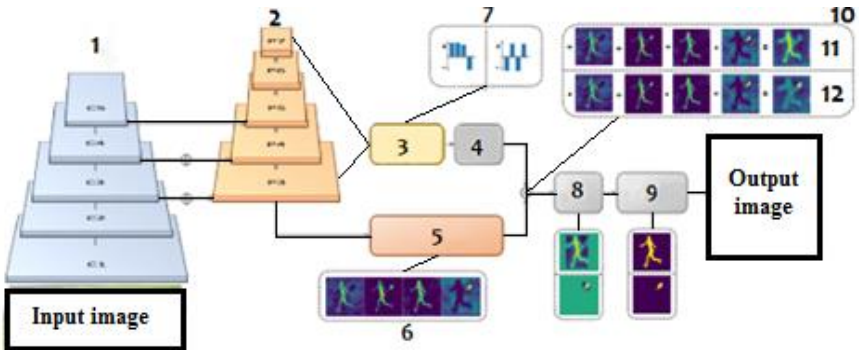


Рис. 1. Архитектура сети YOLACT [1]:
1 – карта признаков; 2 – пирамида признаков; 3 – слои предсказания;
4 – NMS; 5 – протонет; 6 – прототипы; 7 – коэффициенты маски; 8 – обрезка;
9 – порог; 10 – ансамбль; 11, 12 – обнаружение

Fig. 1. YOLACT network architecture [1]:
1 – features map; 2 – features pyramid; 3 – prediction layers;
4 – NMS; 5 – protonet; 6 – prototypes; 7 – mask coefficients;
8 – pruning; 9 – threshold; 10 – ensemble; 11, 12 – detection

Для обучения модели используются три функции потерь (Loss function): Loss классификации ($Lcls$), Loss регрессии рамки вокруг объекта ($Lbox$) и Loss маски ($Lmask$) с весами 1, 1,5 и 6,125 соответственно. Функции $Lcls$ и $Lbox$ определены так же, как в работе [7]. Затем для вычисления $Lmask$ применяется пиксельная двоичная перекрестная энтропия BCE (binary cross entropy) между получившимися масками (M) и масками истинности (Mgt): $Lmask = \text{BCE}\ (M, Mgt)$.

В основу сети положена архитектура сети ResNet-101 (рис. 2) [8] и базовый размер изображения 800×800 пикселов. Для обучения регрессора применяется функция потерь smooth-L_1, для классификации – функция перекрестной энтропии softmax.

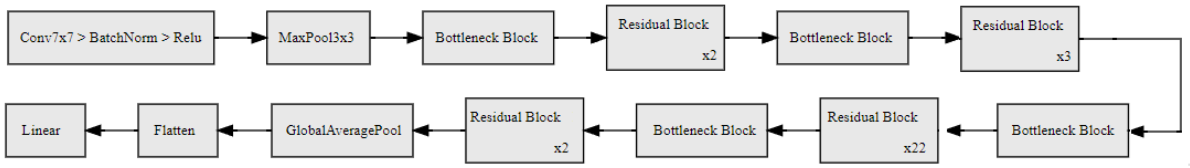


Рис. 2. Архитектура сети ResNet-101

Fig. 2. ResNet-101 network architecture

Модель внимания и ее применение для распознавания объектов электронной коммерции. Модель внимания [2], впервые созданная для машинного перевода, приобрела огромную популярность в сообществе искусственного интеллекта. За прошедшие годы она стала важной частью архитектуры нейронной сети для различных приложений обработки естественного языка, распознавания речи и компьютерного зрения. Модель внимания может интерпретировать нейронные сети и преодолевать ограничения рекуррентных нейронных сетей.

Системы, построенные при помощи модели внимания, фокусируются только на соответствующей части входных данных, полезных для получения необходимых знаний или работы над за-

дачей, и игнорируют несущественные детали. Авторы предлагают модифицировать основу ResNet-101 [8] нейронной сети YOLACT с помощью модели внимания для сосредоточения только на релевантных признаках изображения при решении задачи распознавания изображений товаров электронной коммерции.

Модель внимания фиксирует кросс-канальные корреляции признаков, сохраняя при этом независимое представление в метаструктуре. Модуль сети выполняет набор преобразований для вложений низкой размерности и объединяет их выходные данные. Каждое преобразование включает в себя применение модели внимания по каналам, чтобы зафиксировать взаимозависимости карт признаков, и имеет одну и ту же топологию. Такой подход позволяет ускорить обучение с помощью идентичной реализации, так же как и у унифицированных операторов CNN. Полученный вычислительный блок называется блоком разделения внимания. Объединением нескольких блоков разделения внимания и образуется необходимая архитектура.

Блок разделения внимания состоит из группы карт признаков и операторов разделения внимания. Признаки делятся на группы, которые управляются гиперпараметром мощности (cardinality) группы K. В блок входит также новый гиперпараметр основания R, который отражает количество разделений внутри группы K таким образом, что общее количество групп признаков $G = KR$.

Модель внимания встраивается в сеть ResNet-101 следующим образом: два последовательных полносвязных слоя с числом групп, равным мощности группы, добавляются после объединяющего слоя, чтобы предсказать веса внимания каждого блока. При такой реализации первые сверточные слои 1×1 могут быть объединены в один слой. Сверточные слои 3×3 могут быть реализованы с помощью одной групповой свертки с количеством групп $R \cdot K$. Поэтому блок модели внимания имеет модульную структуру, в которой используются стандартные операторы CNN. Внутренняя структура блока сети ResNet-101 (представлена блоком Bottleneck Block на рис. 2) замещается блоком модели внимания (рис. 3).

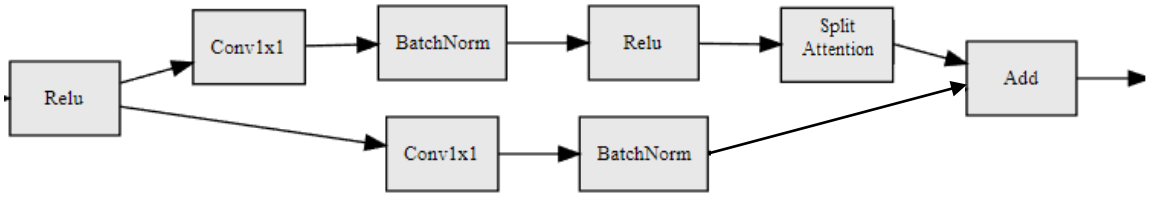


Рис. 3. Архитектура блока сети ResNet-101 с моделью внимания

Fig. 3. ResNet-101 network block architecture with attention model

Как правило, в реализациях ResNet применяется пошаговая свертка на уровне 3×3 вместо уровня 1×1 для сохранения пространственной информации. Сверточные слои требуют обработки границ карты признаков с помощью стратегий заполнения нулями. Этот подход не является оптимальным, поэтому вместо пошаговой свертки в модели внимания используется средний слой пула с размером ядра 3×3.

Модель внимания применялась в слоях основы ResNet-101 совместно с групповой нормализацией, работа выполнялась с использованием фреймворка PyTorch.

Построение обучающего множества. Распознавание изображений товаров электронной коммерции имеет свои особенности по сравнению с распознаванием обычных объектов. Так, например, современные методы распознавания объектов YOLO, SSD, Faster R-CNN и Mask R-CNN оценивают свои алгоритмы на наборах данных PASCAL VOC и MS COCO, в которых распределение данных таково, что более 70 % изображений содержат объекты, принадлежащие к одной категории, а более 50 % – только один экземпляр на изображении, что не соответствует специфике области электронной коммерции. Кроме того, для товаров электронной коммерции характерен большой разброс по качеству снимков. Они могут быть сделаны в профессиональной студии с однородным фоном и освещением, а могут быть размытыми и нечеткими.

Для обучения сети был собран обучающий набор данных, представленный как предварительно обработанными, так и «сырыми» фотографиями продуктов. Он состоит из 62 032 изображений, на которых представлено 26 классов объектов. Набор данных создавался путем сочетания самого идентифицируемого объекта и произвольного фона с использованием методов поворота, растяжения и центрирования.

Каждое изображение имеет размер 800×800 пикселов и хранится в формате RGB, все каналы которого представлены восьмибитной структурой (рис. 4).

Рис. 4. Примеры изображений обучающей выборки

Fig. 4. Examples of images from training dataset

Для решения задачи распознавания каждое изображение из обучающего набора данных также получает метку класса, координаты рамки (bounding box) и бинарную маску (рис. 5) – одноканальное изображение, где 0 обозначает отсутствие объекта или заднего фона, а 255 – наличие объекта или переднего фона.

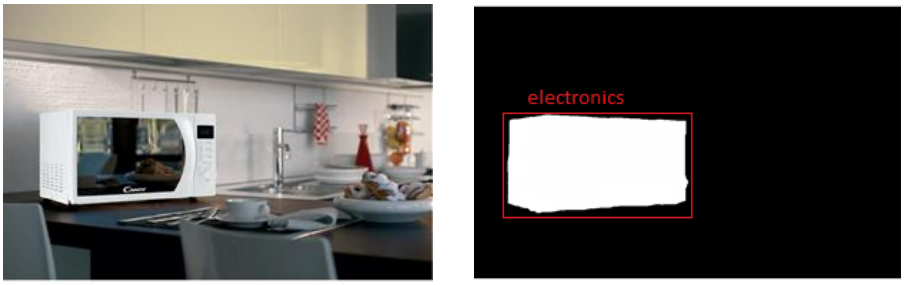


Рис. 5. Оригинальное изображение и его бинарная маска

Fig. 5. Original image and its binary mask

Тестовый набор данных состоит из аналогичных изображений и включает в себя 12 098 изображений. При решении задачи распознавания объектов с помощью нейронных сетей типовым методом является предварительная обработка данных для их стандартизации. Она включает различные алгоритмы в зависимости от поставленной задачи: фильтрация шумов, изменение контраста, подчеркивание границ и т. д.

Для сферы электронной коммерции характерно использование различных по качеству (зашумленности, разрешению, яркости и т. д.) изображений, которые необходимо приводить к единому стандарту – снимкам высокого разрешения со светлым фоном. Для этого изображения сначала необходимо стандартизировать, а затем применить сегментацию для изменения фона.

Поскольку набор данных для обучения ввиду специфики предметной области собирался вручную, при построении обучающего множества для нейронной сети каждое изображение оценива-

лось по следующим параметрам: тусклости, белизне, однородности, размеру и размытости. В случае отклонения по какому-либо признаку изображение проходило процедуру автокоррекции, что не является темой настоящей статьи, поэтому ниже будут показаны результаты работы алгоритмов предобработки данных для изображений, не требующих автокоррекции. Рассмотрим перечисленные параметры более подробно.

Тусклость. Анализ ярких цветов, присутствующих на изображении, помогает определить, тусклое изображение или нет. Данный метод включает в себя следующие шаги:

1) определение всех цветов RGB-изображения;
2) сортировка пикселов изображения;
3) проверка темных тонов (меньше 25 для каждого из RGB-каналов) и их подсчет.

Если итоговый процент больше 85, то изображение считается тусклым. Пример тусклого изображения представлен на рис. 6 (изображение тусклое на 4,93 %).

Рис. 6. Результат работы алгоритма по определению тусклости изображения

Fig. 6. The result of the algorithm for determining the dimness of the image

Белизна. Некоторые изображения могут быть также слишком белыми или яркими. Аналогично определению тусклости (вместо темных пикселов рассматриваются светлые со значением больше 244 по каждому из каналов) производится анализ белизны. Пример работы алгоритма показан на рис. 7 (процент белизны изображения составляет 94,715 %).

Рис. 7. Результат работы алгоритма по определению белизны изображения

Fig. 7. The result of the algorithm for determining the whiteness of the image

Однородность. Некоторые изображения могут не содержать вариаций значений пикселов и быть полностью однородными. Количество перепадов яркости – это мера, которая указывает количество краев, присутствующих на всем изображении. Если это число окажется небольшим, то изображение, скорее всего, является однородным и его сегментация не будет точной. Для построения данного алгоритма использовался метод Кэнни определения границ [9]. Пример однородного изображения показан на рис. 8 (средняя ширина пиксела 0,0998).

Рис. 8. Результат работы алгоритма по определению однородности

Fig. 8. The result of the algorithm for determining the homogeneity

Размытость. Для определения размытости изображения использовалась дисперсия дискретного оператора Лапласа [10]:

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

В этом методе происходит бинаризация изображения, затем производится свертка полученного единственного канала изображения с помощью фильтра. Если указанное значение меньше порогового значения 100, то изображение будет размытым. Пример неразмытого изображения приведен на рис. 9 (коэффициент размытости 157,69087).

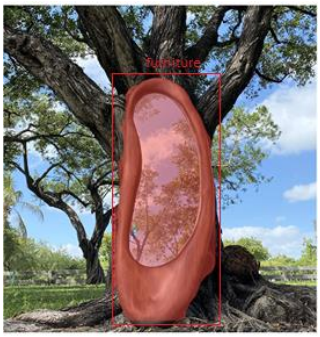
Рис. 9. Результат работы алгоритма по определению размытости

Fig. 9. The result of the algorithm for determining the blur

Результаты распознавания и их обсуждение. В настоящей работе обучались две сети YOLACT: классическая и с применением модели внимания. Обучение модели было направлено на распознавание 26 классов объектов электронной коммерции: модели (человека в полный рост), обуви (четырех классов), одежды (пяти классов), еды (пяти классов), косметики (пяти классов), кухонной техники, аксессуаров и класса заднего фона. Для обучения применялись видеокарта GPU NVIDIA T4 и размер пакета batch_size = 4. Такая величина пакета была выбрана из-за специфики электронной коммерции, так как изображения должны быть высокого разрешения (не менее 800×800 пикселов). Модель внимания использовалась в нейронной сети ResNet-101 для выделения наиболее значимых признаков объекта. Она позволила улучшить распознавание объектов в среднем на 3 %.

Результаты работы классической YOLACT со стандартизацией весов и с применением модели внимания представлены на рис. 10 и в таблице. В качестве метрики использовалась mAP (mean average precision). Из таблицы видно, что добавление механизмов внимания в сеть YOLACT повышает точность предсказания.

a)



b) *c)*

Рис. 10. Результаты работы сети YOLACT: *a)* флаконы; *b)* майка; *c)* зеркало (относится к категории мебели), распознанные сетью

Fig. 10. Results of the YOLACT neural network: a) bottles; b) T-shirt; c) mirror (belongs to furniture category) recognized by the network

Средняя точность работы обученных сетей
Average precision of the trained neural networks

Метод *Method*	Метрики *Metrics*						
	FPS	*AP*	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
YOLACT	28,31	33,7	53,5	35,9	17,2	35,6	45,7
YOLACT со стандартизацией весов [13]	28,31	36,8	59,2	38,2	22,4	37,2	47,2
YOLACT с моделью внимания	28,31	37,4	59,9	39,3	25,2	37,7	48,4

Еще одним результатом применения модели внимания стало увеличение скорости обучения нейронной сети. Для получения результатов, равнозначных полученным стандартной сетью YOLACT, понадобилось в 2,5 раза меньше итераций при обучении.

Обсуждение результатов. В работе предложено использовать модель внимания для задачи распознавания изображений товаров электронной коммерции, что позволит обнаружить корреляции признаков на разных слоях сети для выделения значимой и отсеивания незначимой информации об объекте на изображении.

По итогам проведенного тестирования было выявлено, что обученная сеть обладает свойствами временной стабильности и более быстрой сходимости. Несмотря на это и тот факт, что ре-

зультаты распознавания и сегментации представленной в работе сети более высокого качества, чем у сетей Mask R-CNN [11] и FCIS [12], можно выделить следующие проблемы при их генерации:

1. Ошибка локализации. Если в одном месте на сцене имеется слишком много объектов, сеть может не локализовать каждый объект в собственном прототипе. В этом случае она будет выводить что-то более близкое к маске переднего плана, чем объект, сегментированный по некоторым объектам в группе.

2. Качество данных. Если изображение не соответствует стандартам по однородности, тусклости, размытости и т. д., то стандартные методы автокоррекции не дают ощутимого результата.

Возможным путем решения перечисленных проблем является использование автоматической предобработки данных для улучшения качества изображения.

Добавление модели внимания в архитектуру YOLACT позволяет улучшить как скорость обучения, так и качество распознавания. Эксперименты показали, что улучшение становится более значительным при увеличении сложности сети.

Заключение. В ходе исследования был подготовлен набор данных, произведена его аннотация и построена сеть YOLACT с добавлением модели внимания. Данная модель была обучена на собранном наборе данных, проведены ее валидация и тестирование. Установлено, что модель внимания позволяет сети более точно сосредоточиться на признаках объекта. Это влияет на качество обученной сети и скорость сходимости.

Было выявлено, что использование модели внимания обусловливает значительное улучшение метрик как для распознавания, так и для сегментации объектов. По сравнению с обучением классической сети YOLACT использование модели внимания дает улучшение на 3 % для задачи распознавания.

Предложенный метод может быть применен и к другим архитектурам нейронных сетей, а также использован для YOLACT при увеличении размера входного изображения и, соответственно, выходной маски.

Вклад авторов. *В. В. Сорокина* подготовила набор данных для обучения, произвела его аннотацию, выполнила анализ различных архитектур нейронных сетей, встроила модель внимания в нейронную сеть YOLACT и провела эксперименты по обучению и валидации построенной модели. *С. В. Абламейко* определил цели и задачи исследования, направление работы для достижения поставленных целей, принял участие в анализе и интерпретации результатов.

References

1. Bolya D., Zhou C., Xiao F., Lee Y. J. YOLACT: Real-time instance segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 27 October – 2 November 2019*, pp. 9157–9166.

2. Bahdanau D., Cho K., Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. *3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015*. Available at: https://arxiv.org/abs/1409.0473?context=stat (accessed 01.02.2021).

3. Chaudhuri A., Messina P., Kokkula S., Subramanian A., Krishnan A., …, Kandaswamy V. A smart system for selection of optimal product images in e-commerce. *IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018*, pp. 1728–1736.

4. Zhang X. Content-based e-commerce image classification research. *IEEE Access*, 2020, vol. 8, pp. 160213–160220.

5. Bossard L., Dantone M., Leistner C., Wengert C., Quack T., Van Gool L. Apparel classification with style. *Asian Conference on Computer Vision, Berlin, 2012*, vol. 7727, pp. 321–335.

6. Lao B., Jagadeesh K. Convolutional neural networks for fashion classification and object detection. *CCCV 2015 Computer Vision*, pp. 120–129.

7. Dai J., He K., Li Y., Ren S., Sun J. Instance-sensitive fully convolutional networks. *14th European Conference on Computer Vision, Amsterdam, 11–14 October 2016*, vol. 9910, pp. 534–549.

8. He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016*, 2016, pp. 770–778.

9. Green B. *Canny Edge Detecor*. Available at: https://docs.opencv.org/master/da/d22/tutorial_py_canny.html (accessed 01.02.2021).

10. Pech-Pacheco J. L., Cristobal G., Chamorro-Martinez J., Fernandez-Valdivia J. *Diatom Autofocusing in Brightfield Microscopy: A Comparative Study*. Available at: http://optica.csic.es/papers/icpr2k.pdf (accessed 01.02.2021).

11. He K. Mask R-CNN. *IEEE International Conference on Computer Vision (ICCV), Venice, 22–29 October 2017*, pp. 2980–2988.

12. Qi H., Dai J., Ji X., Wei Y. Fully convolutional instance-aware semantic segmentation. *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, 21–26 July 2017*, pp. 4438–4446.

13. Sorokina V., Ablameyko S. Neural network training acceleration by weight standardization in segmentation of electronic commerce images. *Studies in Computational Intelligence*, 2020, vol. 976, pp. 237–244.

Информация об авторах

Сорокина Виктория Вадимовна, аспирант кафедры веб-технологий и компьютерного моделирования механико-математического факультета, Белорусский государственный университет.
https://orcid.org/0000-0002-2128-1943

Абламейко Сергей Владимирович, академик НАН Беларуси, доктор технических наук, профессор, лауреат Государственной премии Республики Беларусь, заслуженный деятель науки Республики Беларусь, Белорусский государственный университет, Объединенный институт проблем информатики Национальной академии наук Беларуси.
https://orcid.org/0000-0001-9404-1206

Information about the authors

Viktoria V. Sorokina, Postgraduate Student of Web-Technologies and Computer Modeling Department of Mechanics and Mathematics Faculty, Belarusian State University.
https://orcid.org/0000-0002-2128-1943

Sergey V. Ablameyko, Academician of the National Academy of Sciences of Belarus, D. Sc. (Eng.), Professor, Laureate of the State Prize of the Republic of Belarus, Honored Scientist of the Republic of Belarus, Belarusian State University, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus.
https://orcid.org/0000-0001-9404-1206