УДК 004.932.2

МЕТОД ГЕНЕРАЦИИ СИНТЕТИЧЕСКОГО НАБОРА ДАННЫХ ДЛЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

М.М. Лукашевич
докторант БГУИР, руководитель проекта ИООО «Софтек Девелопмент»

А.Н. Макаров
ML инженер ИООО «Софтек Девелопмент»

Н.Ю. Филиппов
студент БГУИР, ML инженер ИООО «Софтек Девелопмент»

Белорусского государственного университет информатики и радиоэлектроники, Республика Беларусь
ИООО «Софтек Девелопмент», Республика Беларусь
E-mail: lukashevich@bsuir.by

М.М. Лукашевич
Является докторантом Белорусского государственного университета информатики и радиоэлектроники. Работает в ИООО «Софтек Девелопмент» в должности руководителя проекта. Сфера научных интересов включает цифровую обработку изображений и распознавание образов, компьютерное зрение.

А.Н. Макаров
Окончил Белорусский государственный университет информатики и радиоэлектроники. Работает в ИООО «Софтек Девелопмент» в должности ML инженера. Основные интересы связаны с решением практических задач в области машинного обучение и компьютерного зрения.

Н.Ю. Филиппов
Является студентом выпускного курса Белорусского государственного университета информатики и радиоэлектроники. Работает в ИООО «Софтек Девелопмент» в должности ML инженера. Основные интересы связаны с решением практических задач в области машинного обучение и глубоких нейронных сетей.

Аннотация. В работе рассмотрены источники данных для задач машинного обучения, некоторые инструменты аннотирования. Особый акцент сделан на применение машинного обучения при решении задач компьютерного зрения. Обоснован подход к созданию синтетических наборов данных. При таком подходе генерируются имитируемые данные для обучения, напоминающие по своим базовым параметрам реальные данные (объекты). Предложен метод генерации синтетических наборов данных, основанный на ряде алгоритмов цифровой обработки изображений таких как, геометрические преобразования, искажения цвета, поворот, добавление шума.

Ключевые слова: машинное обучение, набор данных, разметка данных, синтетические данные.

Введение.

Построение моделей машинного обучения для решения широкого круга задач всегда основывается на данных. Качество моделей машинного обучения напрямую зависит от набора данных, который использовался при обучении модели, его объема и чистоты, вариабельности представленных данных. Условно можно выделить следующие типы данных, с которыми работают в машинном обучение: изображения/видео, табличные

данных, сигналы различной физической природы.

Наборы данных для задач машинного обучения.

Рассмотрим некоторые популярные источники данных. Наиболее известной среди специалистов по машинному обучения является платформа Kaggle [1], предоставляющая доступ к большому числу наборов данных по различным тематикам. Также ресурс содержит большое число задач, их подробное описание, результаты и исходные коды прошедших соревнований, есть возможность загрузить свой набор данных, а также можно воспользоваться вычислительными ресурсами платформы. Ещё одним источником исследовательских данных является Dataset Search от Google [2]. Данные можно найти по таким критериям, как актуальность, формат данных, тип лицензии, тема и др. Здесь представлены данные от различных международных организаций таких как Всемирная организация здравоохранения. Компания Amazon предлагает реестр открытых данных на AWS [3]. Компания Microsoft представляет разработчикам и исследователям доступ к открытым наборам данных Azure [4]. Ресурс содержит данные статистические и научные данные различных организаций. Также имеется возможность создавать собственные облачные базы данных и использовать другие возможности сервиса.

В данной работе мы сконцентрировали свое внимание на данных, представленных в виде изображений. Данные этого типа используются при решении широкого круга задач компьютерного зрения. Особенно популярными наборами изображений для задач компьютерного зрения являются следующие.

База данных ImageNet [5] – это проект по созданию и сопровождения обширнейшей базы аннотированных изображения, предназначенных для обучения и тестирования методов распознавания образов и машинного обучения. Включает в себя более 14 миллионов изображений, организованных в соответствии с иерархией WordNet. Проект сыграл важную роль в продвижении исследований в области компьютерного зрения и глубокого обучения. Данные доступны бесплатно для некоммерческого использования. Примеры данных из базы ImageNet представлены на рисунке 1.

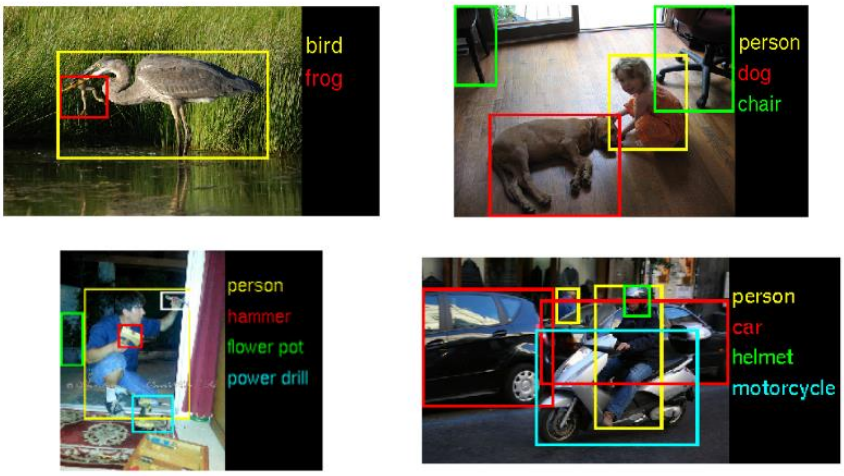

Рисунок 1. Примеры изображений из базы ImageNet [5]

MS COCO [6] – это крупномасштабный набор данных для решения задач обнаружения и сегментации объектов на изображении, аннотирования изображений. Набор содержит 123 287 изображений, на которых представлено 886 284 объектов 22 классов. Примеры данных из базы MS COCO представлены на рисунке 2.

Рисунок 2. Примеры изображений из базы MS COCO [6]

В случае решения задач аннотирования изображений большой интерес представляет набор Visual Genom [7], состоящий из более 100 тысяч аннотированных изображений. Примеры данных из базы Visual Genom представлены на рисунке 3.

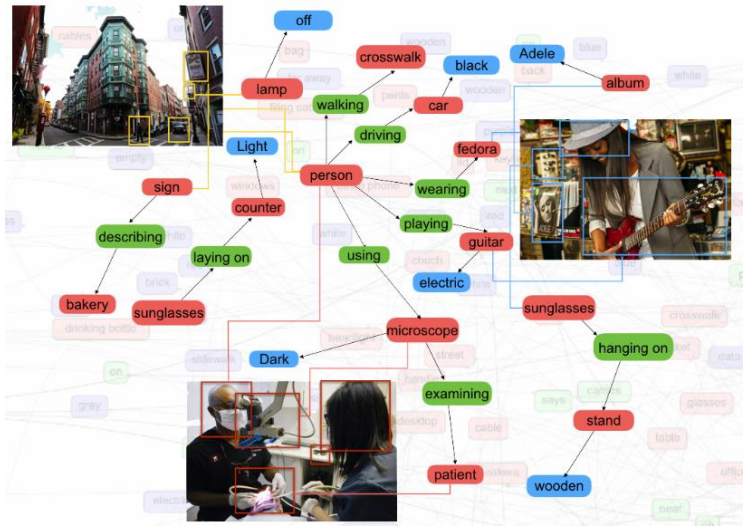

Рисунок 3. Примеры изображений из базы Visual Genom [7]

При разработке приложений, использующих технологии распознавания лиц, стоит обратить внимание на набор данных Labeled Faces in the Wild [8]. Он содержит 3 000 размеченных изображений лиц людей.

Таким образом, в настоящее время для решения различных задач машинного обучения и компьютерного зрения можно найти большое число наборов данных. Одни наборы предназначены для решения широкого круга задач (ImageNet, MS COCO), другие являются узкоспециализированными (Visual Genom). Большинство из них свободно можно использовать лишь в учебных или научных целях. Кроме того, часто они имеют определенные ограничения, например, недостаточный объем данных или ограниченный набор классов. Так же их применение в коммерческих разработках, как правило,

ограничено лицензией.

Подходы к разметке данных.

Отдельно подчеркнем, что набор данных для задач машинного обучения должен содержать не только данные, но метаинформацию. Без размеченных данных нельзя обучить алгоритмы машинного обучения. Данные должны быть интерпретированы и аннотированы в соответствии с решаемой задачей. Также они должны быть чистыми и корректными. Это является значительной частью работы при создании наборов данных.

Самой быстрый и дешевый способ разметки данных – это использование платформ краудсорсинга (например, Toloka.ai [9]). Большая задача по разметке данных разделяется на мелкие и удобные в работе части и доверяется большому числу пользователей, после чего все результаты собираются воедино. Основной проблемой данного подхода является точность разметки данных. В противовес данному решению самой надёжной с точки зрения точности является разметка внутри компании, где обеспечивается полный контроль. Однако не каждая компания может себе позволить содержать штат разметчиков. Одним из решений может быть аусорсинг сторонних специалистов и компаний. Фрилансеры могут быть наняты с помощью таких площадок как UpWork [10]. Но в случае работы с ними нужно иметь четкие инструкции и согласовать используемое программное обеспечение. Привлечение сторонних компаний будет стоить дороже, чем использование фрилансеров, но гарантирует лучшее качество конечного результата. Примерами таких компаний могут быть следующие [11, 12].

Инструменты аннотирования.

При разметке данных могут использовать как собственные инструменты аннотирования, инструменты клиента, так и сторонние инструменты. Собственные инструменты аннотирования могут себе позволить разрабатывать и использовать далеко не все компании. Но это необходимо при решении задач, в которых есть требования к защите данных. Если таких требований нет, то можно воспользоваться программными средствами для аннотирования данных, некоторые из которых являются бесплатными. Назовем несколько наиболее популярных. CVAT [13] - это бесплатный, интерактивный онлайн инструмент аннотирования видео и изображений для компьютерного зрения (рисунок 4). Удобные UI/UX решения данного продукта основаны на отзывах команды профессиональных специалистов по аннотированию данных.

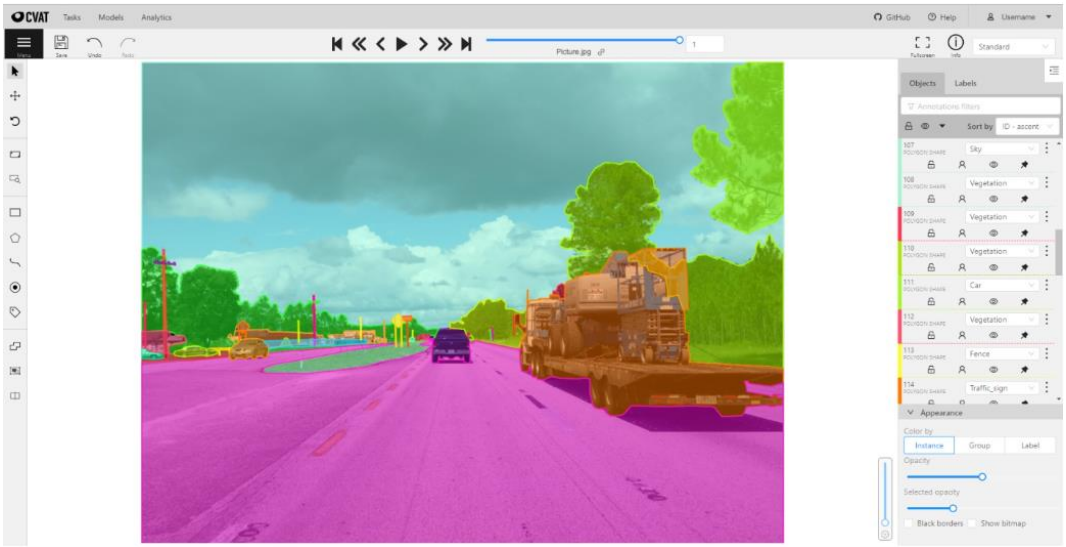

Рисунок 4. CVAT [13]

Еще одним интересным решением является LabelMe [14]. Это графический

инструмент аннотирования изображений, написанный на языке Python и использующий Qt для графического интерфейса.

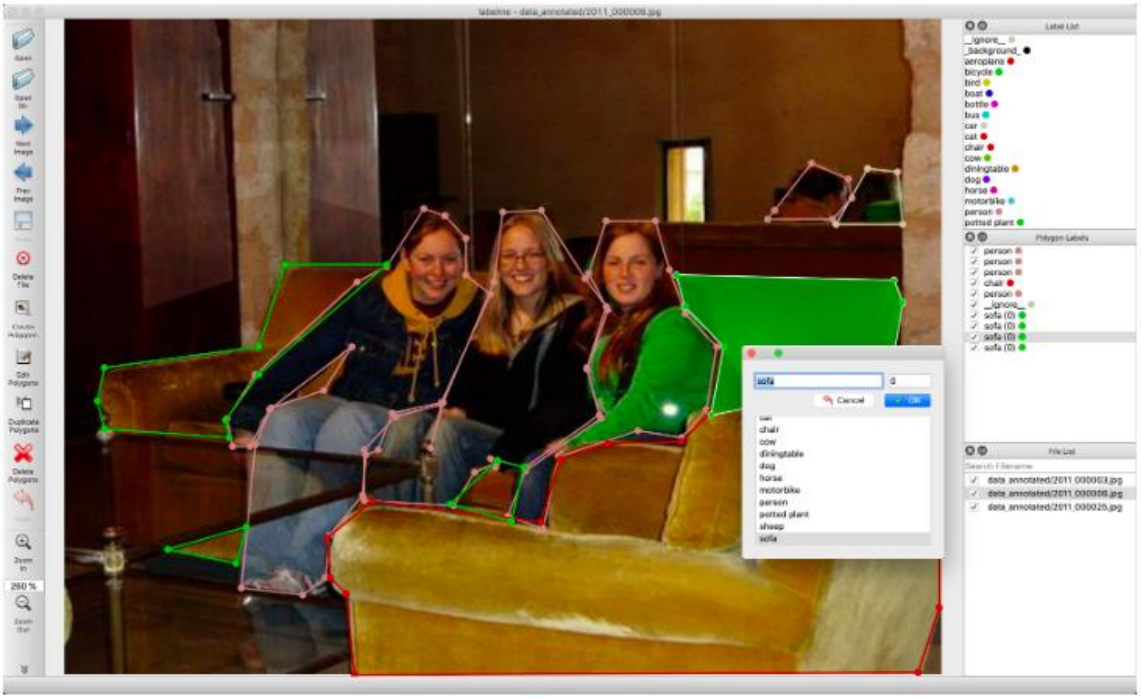

Рисунок 5. CVAT [14]

В связи с выше сказанным актуальной задачей в машинном обучении является задача сбора и разметки наборов данных с целью дальнейшего обучения моделей. При этом зачастую встает вопрос о том, что собрать и разметить большой массив реальных данных (например, изображений) достаточно проблематично в силу большого объема работы, дороговизны, возможной неточности данных и их аннотаций. Отдельным фактором может стать и защита данных. Также в машинном обучение есть задачи, где в принципе нельзя использовать что-либо, кроме синтетических данных. Примеры таких задач – это обучение роботов и беспилотных машин, где используется обучение с подкреплением.

Многие стремятся найти эффективные способы быстрого, качественного и экономного сбора и разметки данных. Чем лучше решение будет предложено, тем быстрее сможет развиваться машинное обучение в частности, и искусственный интеллект в целом.

Метод генерации синтетического набора данных.

Одним из возможных подходов решения данной задачи является создание синтетических наборов данных. Идея заключается в том, что при таком подходе генерируются имитируемые данные для обучения, напоминающие по своим базовым параметрам реальные данные (объекты). Такие данные можно относительно быстро сгенерировать и автоматически разметить. Чаще всего при решении данной задачи в компьютерном зрении используется аугментация - генерация наборов на основе изменения ключевых характеристик данных. Уже есть ряд искусственных наборов данных [15-17] и научных работ по данной тематике [18, 19].

Авторами разработан и предложен метод генерации синтетического набора данных для решения задач машинного обучения. Описанный ниже метод основан на ряде алгоритмов цифровой обработки изображений таких как, геометрические преобразования, искажения цвета, поворот, добавление шума.

Исходными данными при построении набора данных являются 3d модели объектов интереса в формате obj. Формат данных для 3d объекта предполагает представление

объекта в виде набора полигонов (вершин и граней). Вершины задаются через координаты x, y, z и образуют полигон. Они хранятся в нормальной и абсолютной форме. В нормальной форме имеются координаты текстуры, которая представляет набор треугольников.

Введем следующие понятия. Под *объектом интереса* будем понимать объект, принадлежащий какому-либо классу в контексте задач компьютерного зрения (детекция, сегментация, классификация). *Фон* будем считать всю область изображения, за исключением объекта интереса либо изображение, не содержащее объекта интереса. *Нецелевой объект* – это объект, не являющийся объектом интереса, который может являться фоном или содержаться на фоне. *Сцена* – это комбинация фона, целевых и нецелевых объектов.

Метод генерации синтетического набора данных включает в себя следующие шаги:

Шаг 1. Разработка 3d моделей объектов интереса, сбор коллекции моделей. В случае отсутствия возможностей разработки - поиск и сбор коллекции бесплатных 3d моделей.

Шаг 2. Приведение 3d моделей к единому унифицированном виду формате obj. Визуализация моделей с помощью доступных программных инструментов (например, PyOpenGL [20]).

Шаг 3. Для каждой 3d модели выполнение следующих трансформаций (реализация их различных комбинаций):

Шаг 3.1. Цветовая аугментация текстуры.

Шаг 3.2. Масштабные трансформации: изменение расстояния от камеры до объекта.

Шаг 3.3. Изменение угла обзора объекта.

Шаг 4. Обрезка изображения по границам модели.

Шаг 5. Формирование бинарной маски для каждой 3d модели.

Шаг 6. Для каждого фона с целью формирования сцены выполняются следующие операции:

Шаг 6.1. Наложение на фон нецелевых объектов с заданными размерами.

Шаг 6.2. На фон накладываются целевые 2d объекты с различными углами поворота.

Шаг 7. Преобразование сцены: изменение цвета, контраста, яркости, внесение шума.

Шаг 8. Формирование карты фонов и индексов объектов для контроля генерации.

Результаты.

В ходе проведенного эксперимента практическая реализация описанного метода имела следующие особенности. В качестве целевого объекта был взят класс «сумки/рюкзаки/чемоданы», рисунок 6.

Рисунок 6. Примеры изображений целевых объектов

Цветовая аугментация текстуры выполнялась путем перестановки цветовых каналов исходных текстур. Примеры текстур представлены на рисунке 7.

Рисунок 7. Примеры текстур

Анализ показал, что такая трансформация эффективна в случае ярких и насыщенных цветов и бессмысленна в случае градаций серого. Масштабные трансформации выполнялись в 3 вариантах, с сохранением визуального качества объекта при изменении его размера. Изменение угла обзора выполнялось в интервале от -0,5 до 0,5 радиан. Всего было использовано 7 различных значений угла. Число фонов при создании набора данных должно быть достаточно большим. Эксперименты показали, что увеличение числа фонов уменьшает вероятность переобучения модели. Примеры изображений фона представлены на рисунке 8.

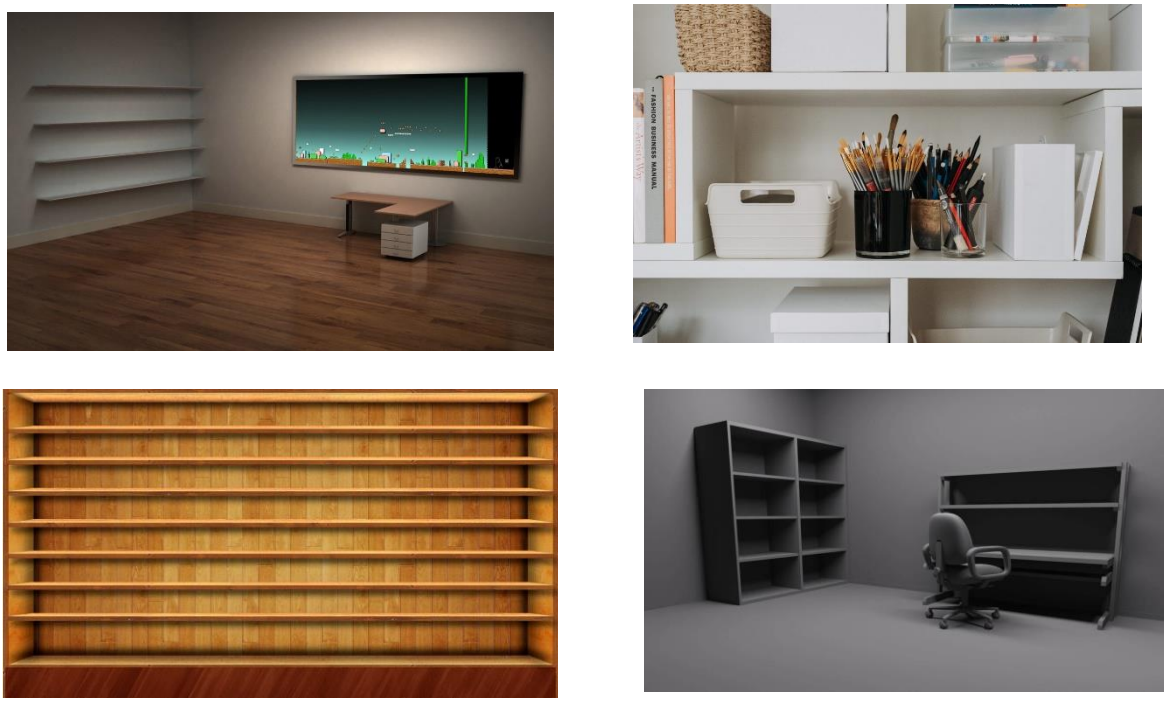

Рисунок 8. Примеры изображений фона

На фон накладывалось не более, чем 10 нецелевых объектов, максимальный размер которых не мог превышать размер фона (1080*1920 пикселей). На фон накладывалось от 4 до 14 целевых объектов с различными углами поворота. Примеры нецелевых объектов

представлены на рисунке 9. В процессе генерации выполнялся контроль пересечения объектов интересов. В ходе экспериментов был получен набор из 2 142 изображений, основой создания которого стали 17 3d моделей.

Рисунок 9. Примеры изображений нецелевых объектов и соответствующих им бинарных масок

Данный набор был использован при решении практической задачи построения модели машинного обучения на основе глубоких нейронных сетей для детекции и трекинга объектов. Модель показала хорошие результаты в условиях отсутствия реальных данных.

Заключение.

Разработанный метод генерации синтетического набора данных может быть использован при создании данных для решения широкого круга задач по компьютерному зрению, где применяется машинное обучение.

Список использованных источников

[1] Kaggle: Your Home for Data Science [Электронный ресурс]. – Режим доступа: https://www.kaggle.com/. – Дата доступа: 24.03.2022.

[2] Dataset Search [Электронный ресурс]. – Режим доступа: https://datasetsearch.research.google.com/. – Дата доступа: 24.03.2022.

[3] Registry of Open Data on AWS [Электронный ресурс]. – Режим доступа: https://registry.opendata.aws/. – Дата доступа: 24.03.2022.

[4] Public data sets for Azure analytics – Azure SQL, Mocrosoft Docs [Электронный ресурс]. – Режим доступа: https://docs.microsoft.com/en-us/azure/azure-sql/public-data-sets. – Дата доступа: 24.03.2022.

[5] ImageNet [Электронный ресурс]. – Режим доступа: https://image-net.org/. – Дата доступа: 24.03.2022.

[6] COCO – Common Objects in Content [Электронный ресурс]. – Режим доступа: https://cocodataset.org/#home. – Дата доступа: 24.03.2022.

[7] VisualGenome [Электронный ресурс]. – Режим доступа: http://visualgenome.org/. – Дата доступа: 24.03.2022.

[8] LFW Face Database : Main [Электронный ресурс]. – Режим доступа: http://vis-www.cs.umass.edu/lfw/. – Дата доступа: 24.03.2022.

[9] Toloka: Data solution to drive AI [Электронный ресурс]. – Режим доступа: https://toloka.ai/. – Дата доступа: 24.03.2022.

[10] Upwork, The World is Work Marketplace[Электронный ресурс]. – Режим доступа: https://www.upwork.com/. – Дата доступа: 24.03.2022.

[11] Artificial Intelligence Companies, Data Science Consulting [Электронный ресурс]. – Режим доступа: https://www.webtunix.com/. – Дата доступа: 24.03.2022.

[12] iMerit: AI Data Soliutions [Электронный ресурс]. – Режим доступа: https://imerit.net/. – Дата доступа: 24.03.2022.

[13] CVTA [Электронный ресурс]. – Режим доступа: https://github.com/openvinotoolkit/cvat. – Дата доступа: 24.03.2022.

[14] LabelMe [Электронный ресурс]. – Режим доступа: https://github.com/wkentaro/labelme. – Дата доступа: 24.03.2022.

[15] Aubry M., Maturana D., Efros A., Russell B., Sivic J. Seeing 3d chairs: exemplar part-based 2d-3d

alignment using a large dataset of cad models — InCVPR, 2014

[16] VC-Clothes [Электронный ресурс]. – Режим доступа: https://wanfb.github.io/dataset.html/.– Дата доступа: 24.03.2022.

[17] Visual Geometry Group - University of Oxford https://www.robots.ox.ac.uk/~vgg/data/scenetext/ — Retrieved January 19, 2021.

[18] SynSys: A Synthetic Data Generation System for Healthcare Applications. Dahmen J, Cook D.Sensors (Basel). 2019 Mar 8;19(5):1181. doi: 10.3390/s19051181.

[19] Sergey I. Nikolenko Synthetic Data for Deep Learning. arXiv:1909.11512 https://doi.org/10.48550/arXiv.1909.11512

[20] PyOpenGL [Электронный ресурс]. – Режим доступа: https://pypi.org/project/PyOpenGL/. – Дата доступа: 24.03.2022.

METHOD FOR SYNTHETIC DATASET GENERATION FOR MACHINE LEARNING TASKS

M.M. LUKASHEVICH
Post Doctoral Researcher of theBelarusian State University of Informatics and Radioelectronics, Unit Lead ML of Softeq Development

A.N. MAKAROV
ML engineer of Softeq Development

N.Y. FILIPPOV
Student of theBelarusian State University of Informatics and Radioelectronics, ML engineer of Softeq Development

Belarusian State University of Informatics and Radioelectronics, Belarus
Softeq Development, Republic of Belarus
E-mail: lukashevich@bsuir.by

Abstract. In this paper the sources of data for machine learning tasks and some annotation tools are discussed. A special emphasis is placed on the application of machine learning to computer vision problems. An approach to generate synthetic datasets is substantiated. This approach generates simulated training data resembling real data (objects) in their basic parameters. A method for generating synthetic datasets based on a number of algorithms for digital image processing, such as geometric transformations, color distortions, rotation, and adding noise is proposed.

Keywords: machine learning, data set, data partitioning, synthetic data.