«Национальный исследовательский университет

«Высшая школа экономики»

Отчет по улучшению группового проекта

**«Классификация изображений,**

**датасет на основе скриншотов/моделей видеоигры»**

Выполнили:

*Баранова Дарья*

*Павлов Сергей*

*Петренко Денис*

*Озеров Максим*

Contents

[Описание датасета 3](#_Toc105367950)

[Промежуточные результаты полученные на первом этапе работы 4](#_Toc105367951)

[Результаты рецензирования 4](#_Toc105367952)

[Первоначальный план улучшения 4](#_Toc105367953)

[Проведенные работы по улучшению решения 5](#_Toc105367954)

[Дальнейшие шаги по улучшению 6](#_Toc105367955)

[Лог метрики качества 7](#_Toc105367956)

[Приложение: Изображения классифицируемых моделей 7](#_Toc105367957)

# Описание датасета

В качестве изображений для классификации были выбраны статические рейнджеры оборудования в космическом корабле игрока из игры «Космические рейнджеры 2: Доминаторы».

Сбор данных был произведён вручную с помощью созданной фанатами утилиты [1], предназначенной для извлечения графических элементов из дистрибутива игры. Такая возможность позволила не пользоваться внутриигровыми скриншотами, которые содержали бы в себе лишние элементы фона и, возможно, были бы хуже исходников по качеству.

В исходниках игры каждое изображение оборудования представляет собой циклический анимированный ряд из ~100 кадров. Поскольку для задачи необходимы статические изображения, для каждой модели оборудования был вручную выбран наиболее репрезентативный (на глаз) кадр.

В итоге каждое исходное изображение обладает следующими характеристиками:

* формат: PNG
* размер: 75 \* 75 pixel
* фон: прозрачный

Исходная выборка содержит 102 изображения, которые относятся к различным 9 классам оборудования со следующим распределением:

|  |  |
| --- | --- |
| Наименование класса | Количество уникальных изображений |
| Оружие (WEAPON) | 15 |
| Генератор защитного поля (DEFGENERATOR) | 11 |
| Двигатель (ENGINE) | 11 |
| Топливный бак (FUELTANKS) | 11 |
| Радар (RADAR) | 11 |
| Ремонтный дроид (REPAIRROBOT) | 11 |
| Сканер (SCANER) | 11 |
| Корпус (HULL) | 11 |
| Захват (CARGOHOOK) | 10 |

Каждый класс представляет собой множество элементов, внешне сильно различающихся между собой. Например, класс «Оружие» состоит из 15 различных видов орудийный инструментов различной формы, размера и цвета. Однако все эти орудийные инструменты принадлежат категории «Оружие», а значит должны быть отнесены к одному классу несмотря на их внешнее сильное различие. Изображения элементов классов представлены в приложении.

Чтобы увеличить объем обучающей выборки исходные изображения были аугментированы разными искажениями до получения 3000 изображений со следующим распределением:

|  |  |
| --- | --- |
| Наименование класса | Количество уникальных изображений |
| Оружие (WEAPON) | 480 |
| Генератор защитного поля (DEFGENERATOR) | 325 |
| Двигатель (ENGINE) | 345 |
| Топливный бак (FUELTANKS) | 322 |
| Радар (RADAR) | 317 |
| Ремонтный дроид (REPAIRROBOT) | 312 |
| Сканер (SCANER) | 312 |
| Корпус (HULL) | 310 |
| Захват (CARGOHOOK) | 277 |

Аугментация проводилась с помощью пайплайна из случайных изменений:

* Поворот (*rotate*) от -8 до +8 градусов с вероятностью 0.5;
* Изменение масштаба (*zoom*) с коэффициентом от 0.8 до 1.01 с вероятностью 0.5;
* Отражение по горизонтали (*flip\_left\_right*) с вероятностью 0.5;
* Случайное искажение (*random\_distortion*) с вероятностью 0.7

# Промежуточные результаты полученные на первом этапе работы

В рамках работы первого этапа были получены следующие результаты:

* Собраны и изучены данные
* Разработано baseline решение задачи. В качестве baseline модели исползовалась SVM. Для изображений проводились различные аугментации. Для извлечения признаков использовалось масштабно-инвариантное преобразование функций SIFT (Scale Invariant Feature Transform). Затем проводилось выделение ключевых точек с помощью кластеризации (K-Means)
* Получены значения метрики качества на тестовой выборке на уровне ~0.88

# Результаты рецензирования

По результатам рецензирования стало понятно, что при разработке пайплайна обучения и проверки модели была допущена утечка данных: в тестовый набор данных попадали аугментированные изображения из трейна. Таким образом достигалась высокое качество метрики, поскольку фактически модель училась на тех же данных, на которых производилась итоговая проверка.

# Первоначальный план улучшения

Первоначальный план улучшений включал следующее шаги:

* Провести подбор гиперпараметров модели;
* Применить другую модель. Например, градиентный бустинг;
* Подобрать на кроссвалидации параметры для модели кластеризации;
* Применить другой подход к кластеризации: dbscan, иерархическую кластеризацию;
* Применить другой алгоритм для извлечения признаков из изображения. Это могут быть следующие алгоритмы:
  1. ORB (Oriented FAST and Rotated BRIEF)
  2. SURF (Speeded Up Robust Features)

# Проведенные работы по улучшению решения

Необходимо отметить, что из-за обнаруженной на этапе рецензирования утечки данных, первоочередной задачей стало её устранение. Устранение утечки данных привело к кардинальному ухудшению метрики качества. Метрика качества baseline решения существенно деградировала. Первоначально оценивалась на уровне ~0.88, а после устранения утечки данных достигала значения порядка ~0.17.

На текущий момент проведены следующее шаги:

1. Применить другую модель. В частности, была применена модель градиентного бустинга в реализации xgboost. Применение модели градиентного бустинга с параметрами по умолчанию дает метрику качества несколько хуже чем метрика откалиброванной модели SVM. Однако факт того, что разница между ними относительно невелика позволил сделать вывод о перспективности дальнейшей более тонкой настройки модели градиентного бустинга.
2. Провести подбор гиперпараметров модели.

Для нахождения параметров модели градиентного бустинга был проведен поиск на сетке параметров из max\_depth, n\_estimators, min\_child\_weights по 5 фолдам. Сетка была построена следующим образом:

max\_depth = range(2, 37, 2)

n\_estimator = [ 8, 16, 24, 32, 40, 48, 56]

min\_child\_weight = [2, 4, 8, 16, 32, 48, 64, 80]

По результатам поиска была найдена оптимальная модель со следующими параметрами:

max\_depth = 4

n\_estimators = 16

min\_child\_weight = 1- (значение по умолчанию)

Полученные значения параметров позволили увеличить значение метрики до 0.22, что значимо лучше значений baseline решения.

Так же во время анализа результатов поиска параметров было обнаружено, что значения метрик качества на тренировочном наборе данных и на тестовом существенно отличаются (см. рисунок ниже, где приведены гистограммы точности). Одной из причин может быть тот факт, что объекты разных классов могут быть очень сильно похожи между собой.

**Chart, histogram

Description automatically generated**

Пример похожих объектов разных классов:

Изображение слева относится к классу «DEFGENERATOR», в то время как все остальные представленные изображения относятся к классу «FUELTANKS»

A picture containing container, glass, indoor

Description automatically generatedA close-up of a robot

Description automatically generated with medium confidenceA picture containing chain

Description automatically generatedA close-up of a robot

Description automatically generated with medium confidenceA close-up of a machine

Description automatically generated with medium confidenceA picture containing stacked

Description automatically generatedA close-up of a factory

Description automatically generated with low confidenceA picture containing engine, automaton

Description automatically generatedA picture containing stacked, displayed

Description automatically generated

1. Также, для модели xgboost была проведена калибровка вероятностей. К сожалению, данная мера не привела к какому-то улучшению метрики качества

# Дальнейшие шаги по улучшению

В первую очередь видится необходимым разобраться со следующими вопросами:

* сильное переобучения модели на тренировочных данных
* похожие объекты, принадлежащие разным классам. Возможно, что решение этого вопроса окажет сильное влияние на пункт выше.

После разрешения указанных выше вопросов должно кажется разумным вернуться к потенциальным улучшениям обозначенным на прошлом этапе работы. А именно:

* Подобрать на кроссвалидации параметры для модели кластеризации;
* Применить другой подход к кластеризации: dbscan, иерархическую кластеризацию;
* Применить другой алгоритм для извлечения признаков из изображения. Это могут быть следующие алгоритмы:
* ORB (Oriented FAST and Rotated BRIEF)
* SURF (Speeded Up Robust Features)

# Лог метрики качества

|  |  |  |
| --- | --- | --- |
| **model** | **test accuracy** | **changes** |
| baseline(svm) | 0,889333 |  |
| baseline(calibrated svm) | 0,934666 |  |
| baseline(svm) | 0,178137 | data leak fixed |
| baseline(calibrated svm) | 0,187584 | data leak fixed |
| xgb | 0,183535 | default params |
| xgb | 0,226720 | model params with grid search |

# Приложение: Изображения классифицируемых моделей

Изображения классифицируемых моделей класса «CARGOHOOK»

A picture containing satellite, automaton

Description automatically generatedA close-up of a logo

Description automatically generated with low confidenceA picture containing chain, metalware

Description automatically generatedA picture containing text

Description automatically generatedA picture containing engine

Description automatically generatedA close up of a toy

Description automatically generated with low confidence

Изображения классифицируемых моделей класса «DEFGENERATOR»

A picture containing lighter, automaton

Description automatically generatedA picture containing container, glass, indoor

Description automatically generatedA close-up of a crystal

Description automatically generated with medium confidenceA picture containing automaton

Description automatically generatedA close-up of a toy

Description automatically generated with medium confidenceA close-up of a computer chip

Description automatically generated with low confidenceA close-up of a machine

Description automatically generated with low confidenceA picture containing engine, automaton, gear

Description automatically generated

Изображения классифицируемых моделей класса «ENGINE»

A picture containing engine, gear

Description automatically generatedA close-up of a fire extinguisher

Description automatically generated with medium confidenceA picture containing transport

Description automatically generatedA picture containing projector, goggles

Description automatically generatedA picture containing skating, roller skating, sport, footwear

Description automatically generated

Изображения классифицируемых моделей класса «FUELTANKS»

A close-up of a machine

Description automatically generated with medium confidenceA close-up of a factory

Description automatically generated with low confidence

Изображения классифицируемых моделей класса «HULL»

A picture containing hydrozoan

Description automatically generatedA picture containing weapon

Description automatically generatedA picture containing weapon

Description automatically generatedA picture containing metalware, gear

Description automatically generatedA picture containing text

Description automatically generatedA picture containing weapon

Description automatically generatedA picture containing frog

Description automatically generatedA picture containing tool

Description automatically generated

Изображения классифицируемых моделей класса «RADAR»

A picture containing microphone

Description automatically generatedA picture containing silver, close

Description automatically generatedA close-up of a glass

Description automatically generated with low confidenceA picture containing stove

Description automatically generatedA close-up of a toy

Description automatically generated with low confidenceA picture containing gear

Description automatically generatedA close-up of a robot

Description automatically generated with low confidenceA picture containing outdoor object, window, ride

Description automatically generated

Изображения классифицируемых моделей класса «REPAIRROBOT»

A picture containing transport, automaton

Description automatically generatedA picture containing automaton

Description automatically generatedA picture containing engine, automaton

Description automatically generatedA picture containing transport, arthropod, satellite

Description automatically generatedA picture containing automaton

Description automatically generated

Изображения классифицируемых моделей класса «SCANER»

A picture containing text, accessory, chain

Description automatically generatedA picture containing person

Description automatically generatedA close-up of a computer

Description automatically generated with low confidenceA picture containing light

Description automatically generatedA close up of a watch

Description automatically generated with low confidenceA close up of a game controller

Description automatically generated with medium confidenceA picture containing text, night

Description automatically generatedA close-up of a light bulb

Description automatically generated with low confidence

Изображения классифицируемых моделей класса «WEAPON»

A close-up of a circuit board

Description automatically generated with low confidenceA close-up of a circuit board

Description automatically generated with low confidenceA picture containing engine

Description automatically generatedA picture containing invertebrate, arthropod, hydrozoan

Description automatically generatedA picture containing satellite

Description automatically generatedA picture containing engine, gear

Description automatically generatedA close up of a fish

Description automatically generated with low confidenceA picture containing engine

Description automatically generatedA close-up of a robot

Description automatically generated with low confidence