

Дипломный проект на тему:
«Classify the morphologies of distant galaxies»

Выполнил: Киселев Александр Константинович

Актуальность темы и ее проблематика:

1. Возникновение Вселенной – один из важнейших фундаментальных вопросов человечества.
2. Классификация и распределение галактик во Вселенной позволяет лучше понять процессы, лежащие в основе их формирования.
3. Более 100 млрд. галактик в видимой части Вселенной.



Рис. 1. Галактика Водоворот и её компаньон NGC 5195

Вводные данные

Работа основана на Kaggle соревновании

Galaxy Zoo - The Galaxy Challenge.

<https://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge>

Совместный проект Kaggle, Galaxy Zoo и Winton Capital.

Вводные данные:

- Тренировочный набор – 61 578 изображений;
- Тестовый набор – 79 975 изображений;
- Размер изображений – 424 x 424 x 3;
- Классификация по 37 признакам.

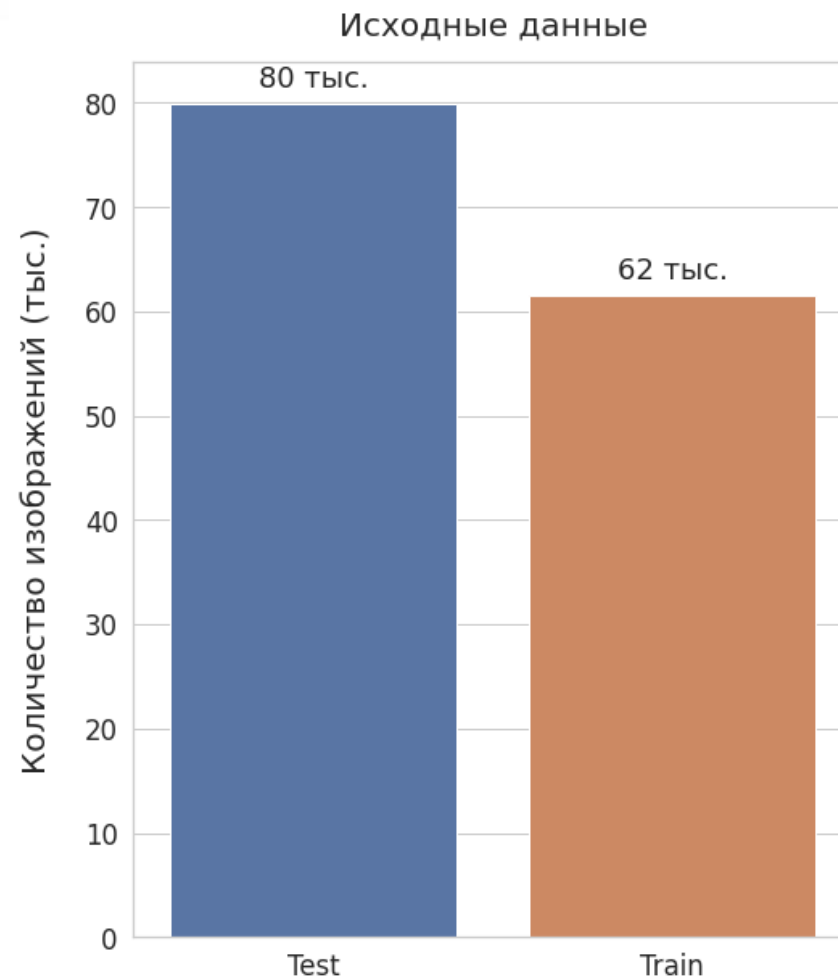


Рис. 1

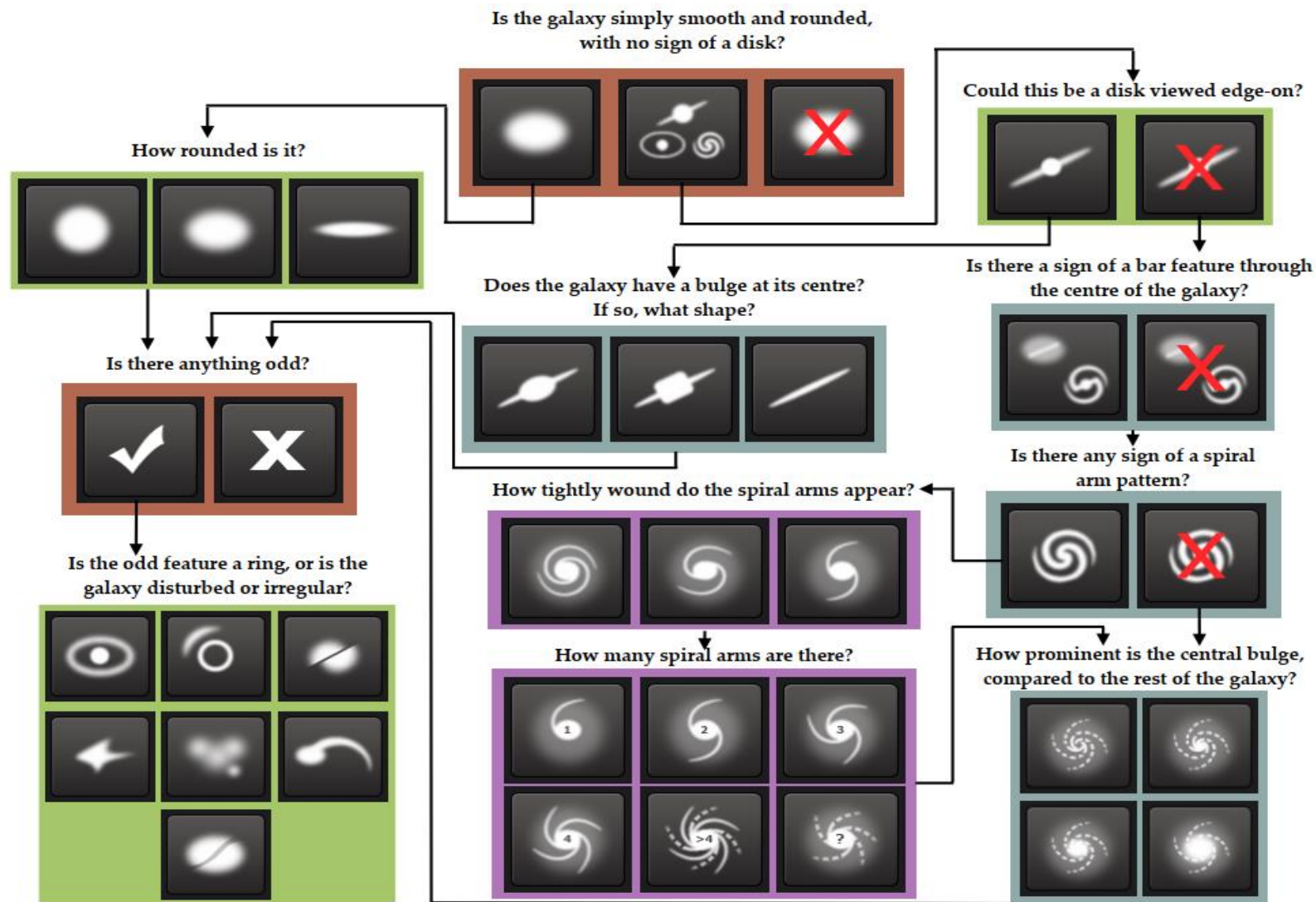


Рис. 3. Классификационное дерево

Тренировочные данные

	GalaxyID	Class1.1	Class1.2	Class1.3	Class2.1	Class2.2	Class3.1	Class3.2	Class4.1	Class4.2	...	Class9.3	Class10.1	Class10.2	Class10.3	Class11.1	Class11.2	Class11.3	Class11.4	Class11.5	Class11.6
0	100008	0.383147	0.616853	0.000000	0.000000	0.616853	0.038452	0.578401	0.418398	0.198455	...	0.000000	0.279952	0.138445	0.000000	0.000000	0.092886	0.000000	0.000000	0.0	0.325512
1	100023	0.327001	0.663777	0.009222	0.031178	0.632599	0.467370	0.165229	0.591328	0.041271	...	0.018764	0.000000	0.131378	0.459950	0.000000	0.591328	0.000000	0.000000	0.0	0.000000
2	100053	0.765717	0.177352	0.056931	0.000000	0.177352	0.000000	0.177352	0.000000	0.177352	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000
3	100078	0.693377	0.238564	0.068059	0.000000	0.238564	0.109493	0.129071	0.189098	0.049466	...	0.000000	0.094549	0.000000	0.094549	0.189098	0.000000	0.000000	0.000000	0.0	0.000000
4	100090	0.933839	0.000000	0.066161	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000
...
61573	999948	0.510379	0.489621	0.000000	0.059207	0.430414	0.000000	0.430414	0.226257	0.204157	...	0.000000	0.226257	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.226257
61574	999950	0.901216	0.098784	0.000000	0.000000	0.098784	0.000000	0.098784	0.000000	0.098784	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000
61575	999958	0.202841	0.777376	0.019783	0.116962	0.660414	0.067245	0.593168	0.140022	0.520391	...	0.000000	0.000000	0.090673	0.049349	0.000000	0.067726	0.000000	0.000000	0.0	0.072296
61576	999964	0.091000	0.909000	0.000000	0.045450	0.863550	0.022452	0.841098	0.795330	0.068220	...	0.000000	0.068398	0.318132	0.408799	0.227464	0.408799	0.090668	0.023065	0.0	0.045334
61577	999967	0.767000	0.140000	0.093000	0.000000	0.140000	0.000000	0.140000	0.023380	0.116620	...	0.000000	0.023380	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.023380

61578 rows × 38 columns

Рис. 4. Решение для тренировочной выборки

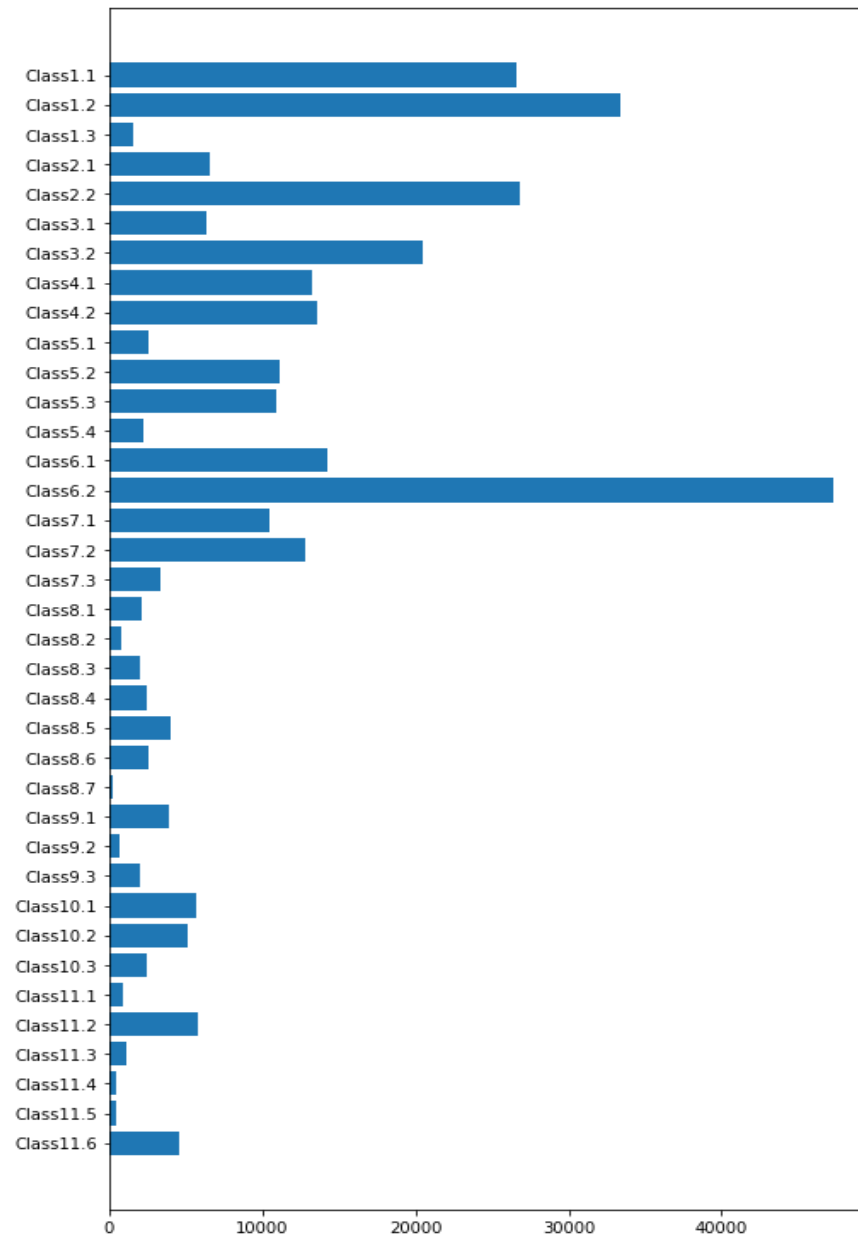
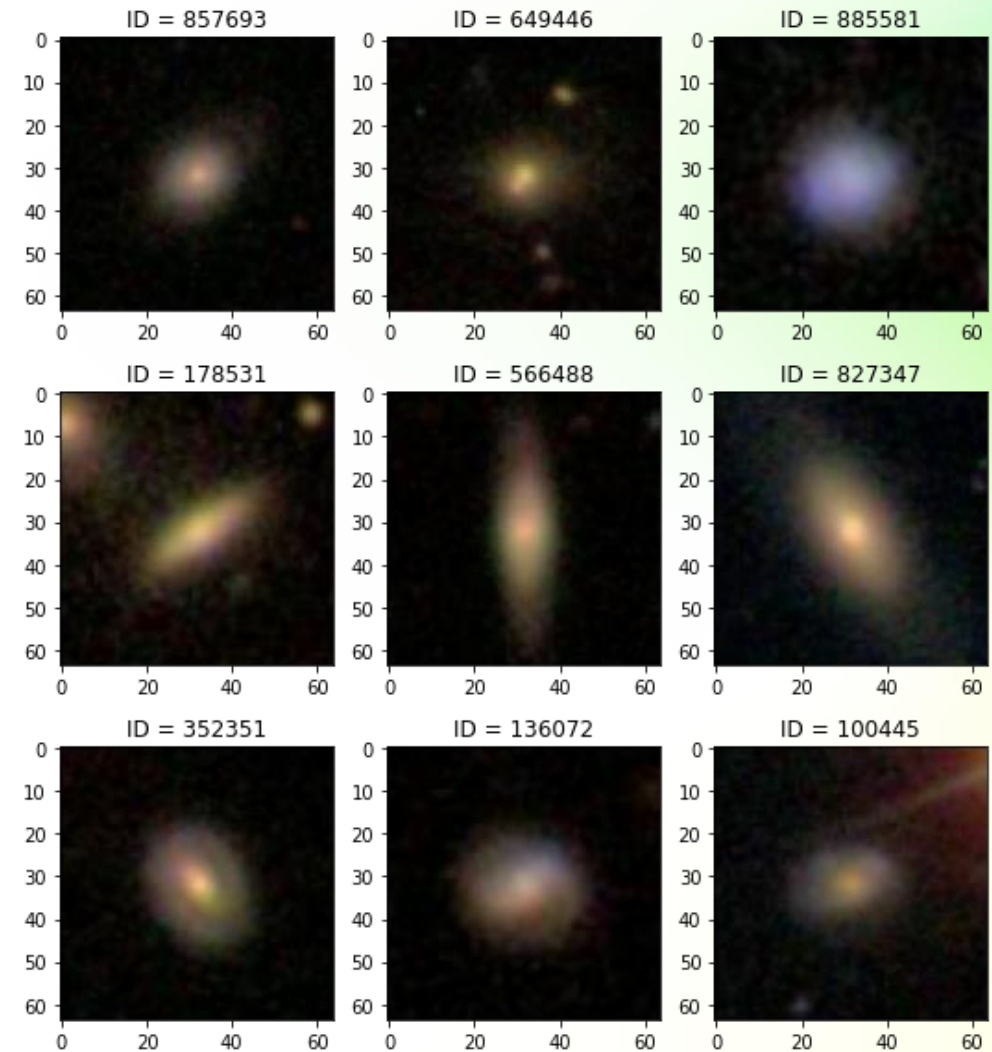
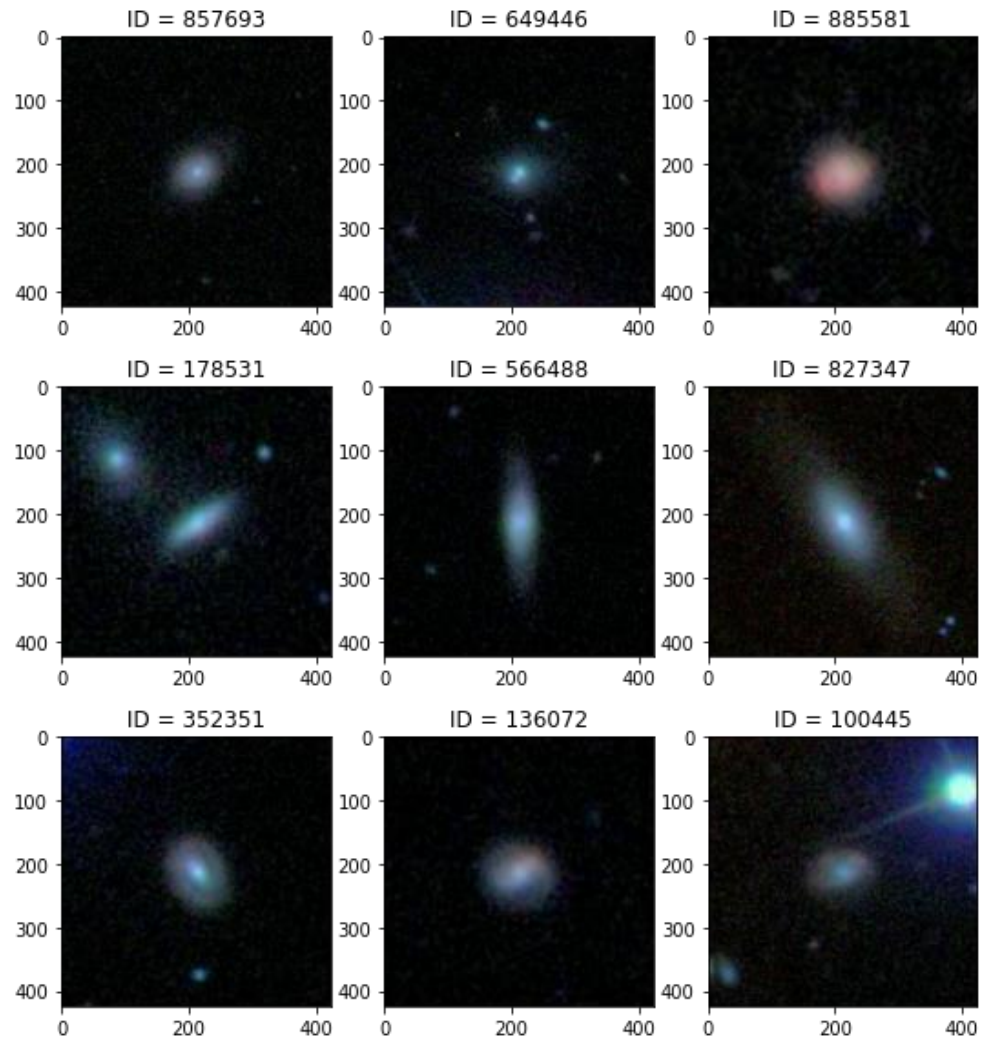


Рис 5. Распределение вероятностей по классам.

Task	Question	Responses	Next
01	<i>Is the galaxy simply smooth and rounded, with no sign of a disk?</i>	smooth features or disk star or artifact	07 02 end
02	<i>Could this be a disk viewed edge-on?</i>	yes no	09 03
03	<i>Is there a sign of a bar feature through the centre of the galaxy?</i>	yes no	04 04
04	<i>Is there any sign of a spiral arm pattern?</i>	yes no	10 05
05	<i>How prominent is the central bulge, compared with the rest of the galaxy?</i>	no bulge just noticeable obvious dominant	06 06 06 06
06	<i>Is there anything odd?</i>	yes no	08 end
07	<i>How rounded is it?</i>	completely round in between cigar-shaped	06 06 06
08	<i>Is the odd feature a ring, or is the galaxy disturbed or irregular?</i>	ring lens or arc disturbed irregular other merger dust lane	end end end end end end end
09	<i>Does the galaxy have a bulge at its centre? If so, what shape?</i>	rounded boxy no bulge	06 06 06
10	<i>How tightly wound do the spiral arms appear?</i>	tight medium loose	11 11 11
11	<i>How many spiral arms are there?</i>	1 2 3 4 more than four can't tell	05 05 05 05 05 05

Рис. 6. Описание классов

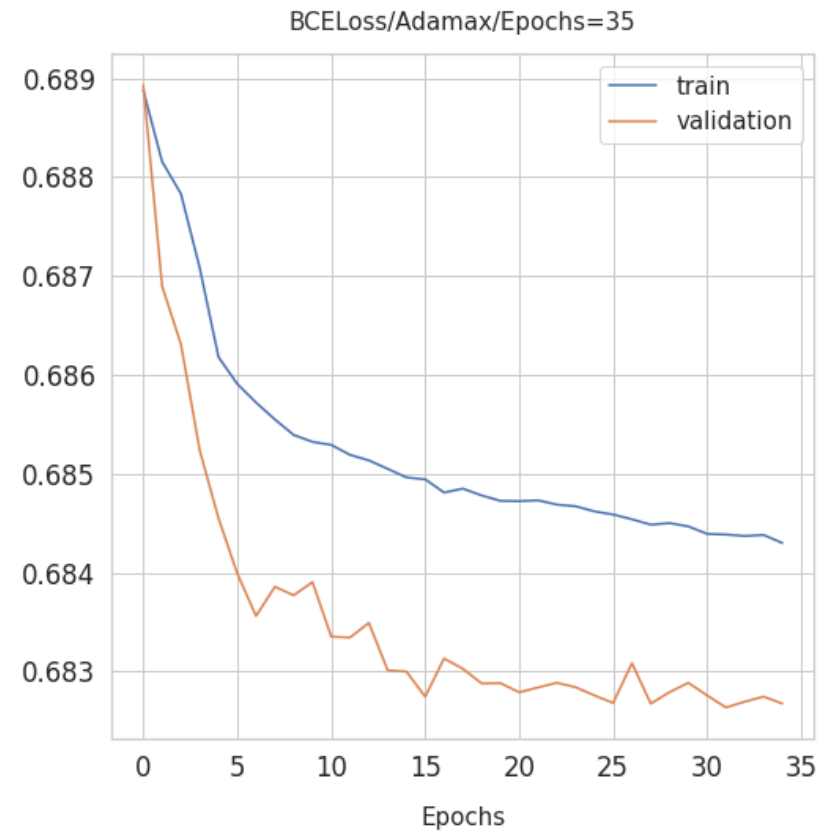
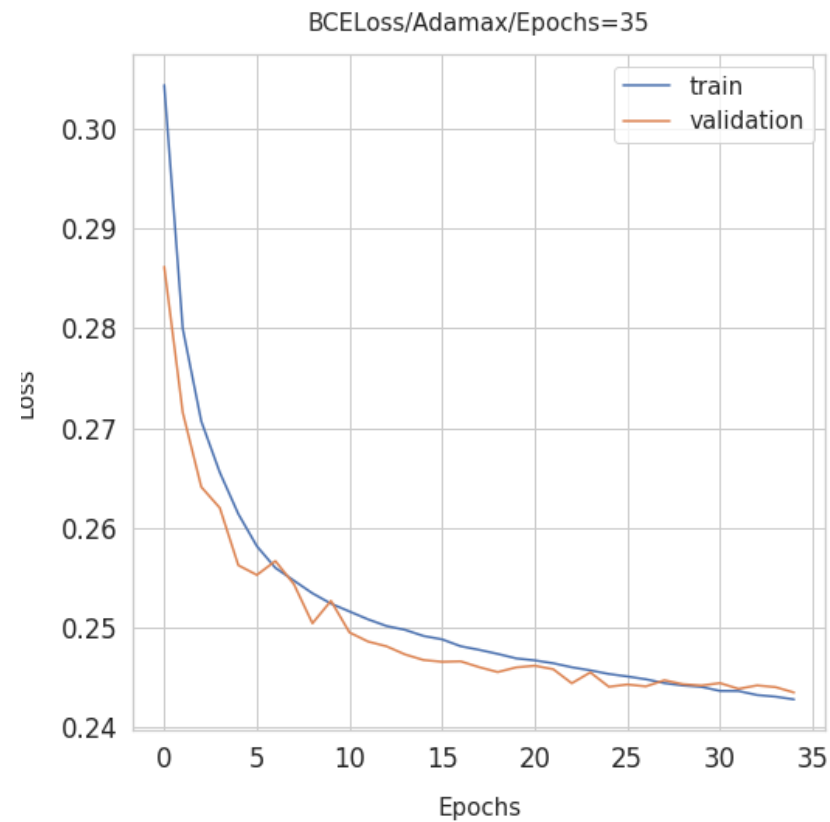
Уменьшение размера изображений



Модель

Два варианта свёрточной нейронной сети.

Оптимизатор – Adamax, функция потерь – BCELoss, количество обучающих эпох – 35.



Выводы

Private Score

Public Score

0.10215

0.10254

53-е место в общем зачете Kaggle. Результат достигнут с использованием одной нейросети. При увеличении количества эпох возможно было бы получить большую точность.

Лучшие результаты в соревновании достигнуты путем объединения результатов нескольких моделей, с использованием аугментации и других методов работы с изображениями.

Лучший результат показали оптимизаторы Adam и Adamax.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Дж. Вандер Плас. Python для сложных задач
2. <https://pytorch.org/>
3. <https://scikit-learn.org/>
4. <https://numpy.org/>
5. <https://pandas.pydata.org/>
6. <https://partner.sberbank-school.ru/>
7. <https://habr.com/>

Спасибо за внимание!