Королихин Владимир, 471 группа

Задача

Необходимо было расширить существующий классификатор тРНК на археи, бактерии, растения и грибы на уровень вглубь по дереву жизни -- научиться различать между собой некоторые основные типы бактерий.

Подготовка данных

Данные собирались из базы https://rnacentral.org/.

Для класса other выбирались все данные по запросу bacteria, но не включающие 8 классов, уже участвующих в классификации. Дальше нужное количество цепочек выбиралось случайным образом.

В работе исследовались датасеты двух видов: с классом other и без.
Для восьми классов бактерий использовались данные в следующем соотношении: train/valid/test = 32000/4000/6000

Датасет, включающий other дополнительно содержал train/valid/test = 12000/1500/750

Классы Fusobacteria и Chlamidiae имеют дубликаты (по 3 на каждую цепочку), так как цепочек в базе по данным классам в разы меньше.

Изменения модели

Для решения поставленной задачи использовалась модель на основе “imagesExtended” и проводились эксперименты по изменению последних слоев, а также самих параметров модели.

Эксперименты с моделями, классифицирующими на 9 классов

1. Модель со слоями после слоя 30420: 4096-1024-256 (lr = 0.05, размер батча 32)

	precision	recall	f1-score	support
Actinobacteria	0.67	0.64	0.66	750
Bacteroidetes	0.82	0.59	0.69	750
Chlamidiae	0.89	0.56	0.69	750
Firmicutes	0.64	0.59	0.62	750
Fusobacteria	0.96	0.67	0.79	750
Proteobacteria	0.81	0.58	0.68	750
Spirochaetes	0.83	0.61	0.70	750
Tenericutes	0.81	0.63	0.71	750
other	0.28	0.79	0.41	750
accuracy			0.63	6750
macro avg	0.75	0.63	0.66	6750
weighted avg	0.75	0.63	0.66	6750

2. Модель со слоями после слоя 30420: 4096-1024-256 (lr = 0.06, размер батча 32)

	precision	recall	f1-score	support
Actinobacteria	0.70	0.70	0.70	750
Bacteroidetes	0.84	0.56	0.67	750
Chlamidiae	0.84	0.57	0.68	750
Firmicutes	0.68	0.58	0.63	750
Fusobacteria	0.96	0.70	0.81	750
Proteobacteria	0.77	0.63	0.69	750
Spirochaetes	0.86	0.58	0.69	750
Tenericutes	0.82	0.63	0.72	750
other	0.28	0.79	0.42	750
accuracy			0.64	6750
macro avg	0.75	0.64	0.67	6750
weighted avg	0.75	0.64	0.67	6750

3. Модель со слоями после слоя 30420: 4096-1024-256 (lr = 0.07, размер батча 32)

	precision	recall	f1-score	support
Actinobacteria	0.70	0.68	0.69	750
Bacteroidetes	0.80	0.64	0.71	750
Chlamidiae	0.87	0.62	0.72	750
Firmicutes	0.69	0.58	0.63	750
Fusobacteria	0.94	0.71	0.81	750
Proteobacteria	0.75	0.66	0.70	750
Spirochaetes	0.84	0.61	0.71	750
Tenericutes	0.81	0.68	0.74	750
other	0.31	0.76	0.44	750
accuracy			0.66	6750
macro avg	0.75	0.66	0.68	6750
weighted avg	0.75	0.66	0.68	6750

4. Модель со слоями после слоя 30420: 4096-1024-256 (lr = 0.08, размер батча 32)

	precision	recall	f1-score	support
Actinobacteria	0.69	0.71	0.70	750
Bacteroidetes	0.85	0.60	0.70	750
Chlamidiae	0.86	0.59	0.70	750
Firmicutes	0.70	0.62	0.66	750
Fusobacteria	0.97	0.71	0.82	750
Proteobacteria	0.80	0.63	0.71	750
Spirochaetes	0.83	0.64	0.73	750
Tenericutes	0.82	0.68	0.74	750
other	0.31	0.79	0.44	750
accuracy			0.66	6750
macro avg	0.76	0.66	0.69	6750
weighted avg	0.76	0.66	0.69	6750

5. Модель со слоями после слоя 30420: 4096-1024-256 (lr = 0.07 с автопонижением lr каждые 25 эпох, размер батча 32)

	precision	recall	f1-score	support
Actinobacteria	0.69	0.71	0.70	750
Bacteroidetes	0.85	0.60	0.70	750
Chlamidiae	0.86	0.59	0.70	750
Firmicutes	0.70	0.62	0.66	750
Fusobacteria	0.97	0.71	0.82	750
Proteobacteria	0.80	0.63	0.71	750
Spirochaetes	0.83	0.64	0.73	750
Tenericutes	0.82	0.68	0.74	750
other	0.31	0.79	0.44	750
accuracy			0.66	6750
macro avg	0.76	0.66	0.69	6750
weighted avg	0.76	0.66	0.69	6750

6. Модель со слоями после слоя 30420: 4096-2048-1024-256 (lr = 0.05, размер батча 32)

	precision	recall	f1-score	support
Actinobacteria	0.70	0.71	0.71	750
Bacteroidetes	0.84	0.54	0.66	750
Chlamidiae	0.87	0.56	0.68	750
Firmicutes	0.64	0.61	0.62	750
Fusobacteria	0.96	0.69	0.81	750
Proteobacteria	0.81	0.60	0.69	750
Spirochaetes	0.80	0.61	0.69	750
Tenericutes	0.83	0.63	0.72	750
other	0.28	0.78	0.41	750
accuracy			0.64	6750
macro avg	0.75	0.64	0.67	6750
weighted avg	0.75	0.64	0.67	6750

7. Модель со слоями после слоя 30420: 4096-1024-256 (lr = 0.05, замороженные верхние слои, размер батча 32)

	precision	recall	f1-score	support
Actinobacteria	0.39	0.24	0.29	750
Bacteroidetes	0.55	0.03	0.06	750
Chlamidiae	0.79	0.09	0.16	750
Firmicutes	0.34	0.36	0.35	750
Fusobacteria	0.88	0.16	0.27	750
Proteobacteria	0.50	0.39	0.44	750
Spirochaetes	0.58	0.05	0.10	750
Tenericutes	0.89	0.10	0.17	750
other	0.14	0.83	0.24	750
accuracy			0.25	6750
macro avg	0.56	0.25	0.23	6750
weighted avg	0.56	0.25	0.23	6750

8. Модель со слоями после слоя 30420: 4096-1024-256 (lr=0.07, размер батча 64)

	precision	recall	f1-score	support
Actinobacteria	0.74	0.70	0.72	750
Bacteroidetes	0.82	0.63	0.71	750
Chlamidiae	0.87	0.61	0.71	750
Firmicutes	0.69	0.62	0.65	750
Fusobacteria	0.97	0.71	0.82	750
Proteobacteria	0.81	0.68	0.74	750
Spirochaetes	0.85	0.63	0.72	750
Tenericutes	0.83	0.75	0.79	750
other	0.32	0.78	0.45	750
accuracy			0.68	6750
macro avg	0.77	0.68	0.70	6750
weighted avg	0.77	0.68	0.70	6750

9. Модель со слоями после слоя 30420: 4096-2048-1024-256-128 (lr = 0.05, размер батча 32)

	precision	recall	f1-score	support
Actinobacteria	0.75	0.65	0.70	750
Bacteroidetes	0.86	0.54	0.66	750
Chlamidiae	0.88	0.55	0.68	750
Firmicutes	0.56	0.54	0.55	750
Fusobacteria	0.97	0.67	0.79	750
Proteobacteria	0.74	0.52	0.61	750
Spirochaetes	0.79	0.59	0.68	750
Tenericutes	0.81	0.60	0.69	750
other	0.26	0.78	0.39	750
accuracy			0.61	6750
macro avg	0.74	0.61	0.64	6750
weighted avg	0.74	0.61	0.64	6750

10. Модель со слоями после слоя 30420: 2048-4096-8192-2048-1024-128 (lr = 0.05, размер батча 32)

	precision	recall	f1-score	support
Actinobacteria	0.69	0.69	0.69	750
Bacteroidetes	0.79	0.60	0.68	750
Chlamidiae	0.84	0.56	0.67	750
Firmicutes	0.40	0.46	0.43	750
Fusobacteria	0.97	0.69	0.80	750
Proteobacteria	0.72	0.43	0.54	750
Spirochaetes	0.82	0.58	0.68	750
Tenericutes	0.82	0.61	0.70	750
other	0.28	0.73	0.40	750
accuracy			0.60	6750
macro avg	0.70	0.60	0.62	6750
weighted avg	0.70	0.60	0.62	6750

Эксперименты с моделями, классифицирующими на 8 классов (без other)

1. Модель со слоями после слоя 30420: 4096-1024-256 (lr = 0.05, размер батча 32)

	precision	recall	f1-score	support
Actinobacteria	0.52	0.86	0.65	750
Bacteroidetes	0.81	0.69	0.74	750
Chlamidiae	0.81	0.66	0.73	750
Firmicutes	0.56	0.74	0.64	750
Fusobacteria	0.94	0.73	0.82	750
Proteobacteria	0.79	0.69	0.74	750
Spirochaetes	0.77	0.71	0.74	750
Tenericutes	0.83	0.66	0.73	750
accuracy			0.72	6000
macro avg	0.75	0.72	0.72	6000
weighted avg	0.75	0.72	0.72	6000

2. Модель со слоями после слоя 30420: 4096-1024-256 (lr = 0.07, размер батча 32)

	precision	recall	f1-score	support
Actinobacteria	0.57	0.86	0.69	750
Bacteroidetes	0.78	0.71	0.75	750
Chlamidiae	0.83	0.65	0.73	750
Firmicutes	0.57	0.72	0.64	750
Fusobacteria	0.95	0.74	0.83	750
Proteobacteria	0.74	0.74	0.74	750
Spirochaetes	0.74	0.74	0.74	750
Tenericutes	0.82	0.65	0.73	750
accuracy			0.72	6000
macro avg	0.75	0.72	0.73	6000
weighted avg	0.75	0.72	0.73	6000

3. Модель со слоями после слоя 30420: 4096-1024-256 (lr = 0.08, размер батча 32)

	precision	recall	f1-score	support
Actinobacteria	0.59	0.84	0.69	750
Bacteroidetes	0.77	0.72	0.75	750
Chlamidiae	0.82	0.65	0.72	750
Firmicutes	0.57	0.77	0.66	750
Fusobacteria	0.95	0.73	0.82	750
Proteobacteria	0.75	0.71	0.73	750
Spirochaetes	0.77	0.71	0.74	750
Tenericutes	0.81	0.71	0.76	750
accuracy			0.73	6000
macro avg	0.75	0.73	0.73	6000
weighted avg	0.75	0.73	0.73	6000

4. Модель со слоями после слоя 30420: 4096-1024-256 (lr = 0.07 с автопонижением lr каждые 25 эпох, размер батча 32)

	precision	recall	f1-score	support
Actinobacteria	0.58	0.85	0.69	750
Bacteroidetes	0.79	0.68	0.73	750
Chlamidiae	0.82	0.64	0.72	750
Firmicutes	0.57	0.74	0.64	750
Fusobacteria	0.95	0.73	0.82	750
Proteobacteria	0.70	0.73	0.72	750
Spirochaetes	0.76	0.71	0.73	750
Tenericutes	0.83	0.68	0.75	750
accuracy			0.72	6000
macro avg	0.75	0.72	0.72	6000
weighted avg	0.75	0.72	0.72	6000

5. Модель со слоями после слоя 30420: 4096-2048-1024-256 (lr = 0.05, размер батча 32)

	precision	recall	f1-score	support
Actinobacteria	0.53	0.87	0.66	750
Bacteroidetes	0.80	0.67	0.73	750
Chlamidiae	0.83	0.63	0.72	750
Firmicutes	0.59	0.69	0.63	750
Fusobacteria	0.92	0.72	0.81	750
Proteobacteria	0.72	0.73	0.72	750
Spirochaetes	0.72	0.71	0.72	750
Tenericutes	0.84	0.68	0.75	750
accuracy			0.71	6000
macro avg	0.74	0.71	0.72	6000
weighted avg	0.74	0.71	0.72	6000

6. Модель со слоями после слоя 30420: 4096-1024-256 (lr = 0.05, замороженные верхние слои, размер батча 32)

	precision	recall	f1-score	support
Actinobacteria	0.18	0.54	0.27	750
Bacteroidetes	0.59	0.10	0.17	750
Chlamidiae	0.77	0.12	0.21	750
Firmicutes	0.25	0.79	0.38	750
Fusobacteria	0.85	0.14	0.24	750
Proteobacteria	0.47	0.51	0.49	750
Spirochaetes	0.48	0.06	0.11	750
Tenericutes	0.85	0.10	0.17	750
accuracy			0.29	6000
macro avg	0.55	0.29	0.26	6000
weighted avg	0.55	0.29	0.26	6000

7. Модель со слоями после слоя 30420: 4096-1024-256 (lr=0.07, размер батча 64)

	precision	recall	f1-score	support
Actinobacteria	0.61	0.86	0.71	750
Bacteroidetes	0.80	0.70	0.75	750
Chlamidiae	0.84	0.63	0.72	750
Firmicutes	0.57	0.75	0.65	750
Fusobacteria	0.98	0.69	0.81	750
Proteobacteria	0.70	0.77	0.74	750
Spirochaetes	0.75	0.74	0.74	750
Tenericutes	0.81	0.68	0.74	750
accuracy			0.73	6000
macro avg	0.76	0.73	0.73	6000
weighted avg	0.76	0.73	0.73	6000

8. Модель со слоями после слоя 30420: 2048-4096-8192-2048-1024-128 (lr = 0.05, размер батча 32)

	precision	recall	f1-score	support
Actinobacteria	0.53	0.85	0.65	750
Bacteroidetes	0.78	0.66	0.72	750
Chlamidiae	0.84	0.57	0.68	750
Firmicutes	0.50	0.73	0.59	750
Fusobacteria	0.93	0.72	0.81	750
Proteobacteria	0.67	0.74	0.70	750
Spirochaetes	0.79	0.62	0.69	750
Tenericutes	0.79	0.61	0.69	750
accuracy			0.69	6000
macro avg	0.73	0.69	0.69	6000
weighted avg	0.73	0.69	0.69	6000

Выводы

- Наиболее оптимальными являются lr = 0.06-0.07
- Добавление больше трех новых слоев вызывает сильное переобучение
- Во всех моделях наблюдается переобучение, а увеличение `Dropout` приводит ухудшению `accuracy`
- Заморозка части слоев значительно ухудшает accuracy
- Увеличение размера батча (до 64) позитивно сказывается на модели
- Внедрение динамического изменения lr не влияет на качество обучения
- Сильное сужение на предпоследнем слое ухудшает результат (256 оптимально)
- Увеличение количества эпох (> 150) не улучшает результат

Результат обучения:

- Для данных с other: лучшее accuracy 0.68 (модель 8)
- Для данных без other лучшее accuracy 0.73 (модель 7)