

Классификация паттернов белков (Kaggle Human Protein Atlas Image Classification)

Дмитрий Буслов, SAP

Наши результаты

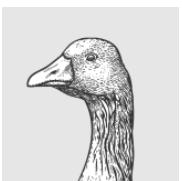
Команда:



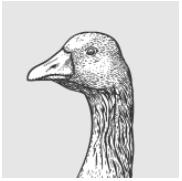
Sergei Fironov



Dmitry Buslov



Dmytro Panchenko



Alexander Kiselev

Public

#	$\Delta 1w$	Team Name	Kernel	Team Members	Score	Entries
1	▲ 14	bestfitting			0.65602	130
2	▼ 1	Arm			0.65585	283
3	▲ 1	Wienerschnitzelgemeinschaft			0.65331	498
4	▼ 2	One More Layer (Of Stacking)			0.64799	269

Private

#	Δ_{pub}	Team Name	Kernel	Team Members	Score	Entries
1	—	bestfitting			0.59369	130
2	▲ 4	WAIR			0.57152	234
3	▲ 14	pudae			0.57008	115
4	▼ 1	Wienerschnitzelgemeinschaft			0.56766	498
5	—	vpp			0.56696	334
6	▲ 7	YaG320			0.56621	105
7	▲ 3	Guanshuo Xu			0.56335	140
8	▼ 4	One More Layer (Of Stacking)			0.56309	269

Описание задачи

- Классификация протеинов (Multilabel)
- 28 классов
- 31.1k train, 11.7k test
- Метрика - F1-macro
- 4 канала (RGBY)

```
# proj_kaggle_protein
```



n01z3 9:22 PM



Я создал



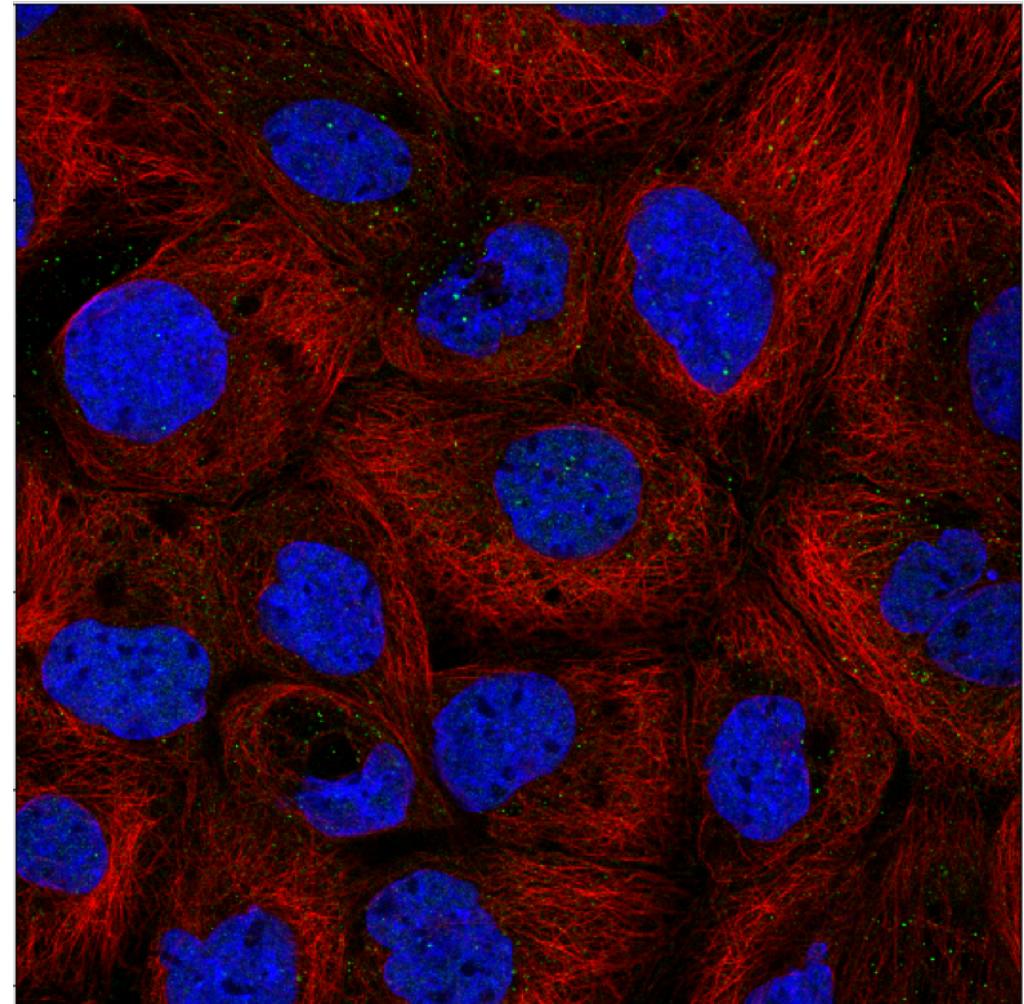
31



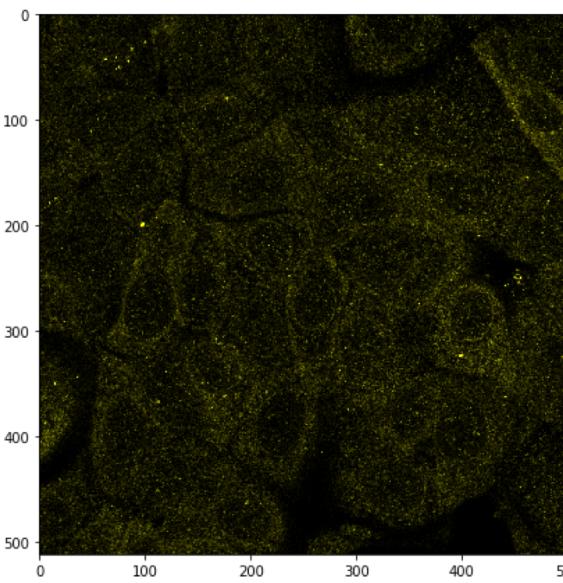
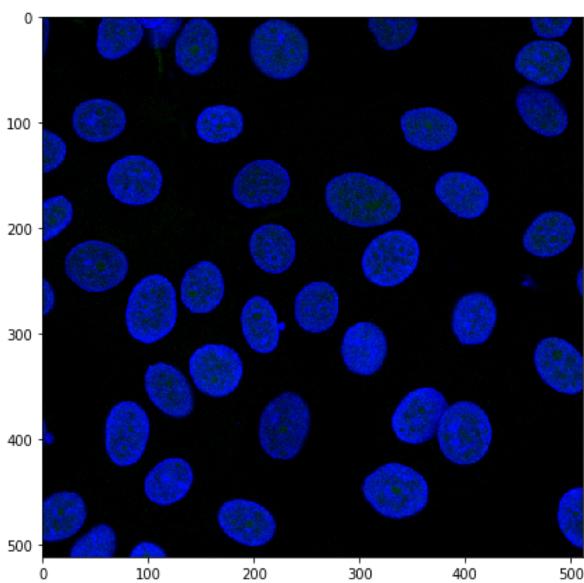
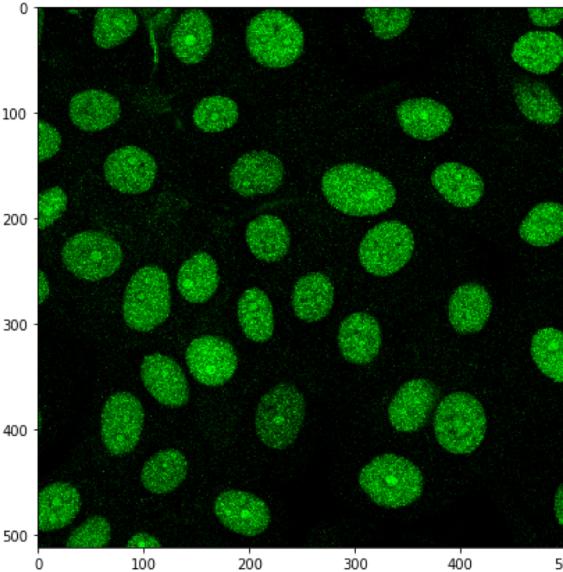
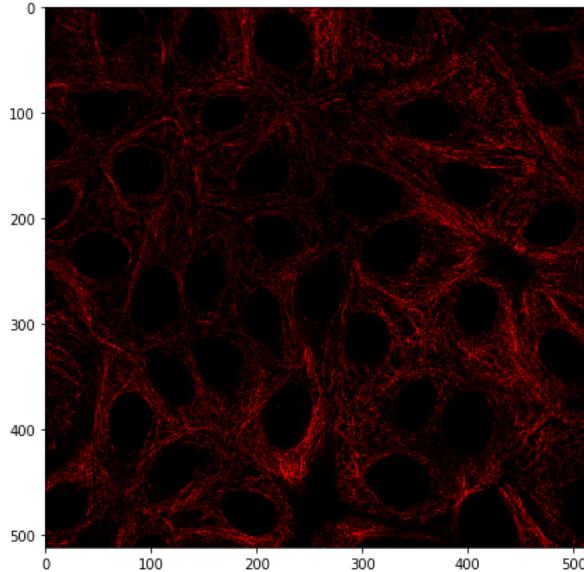
5



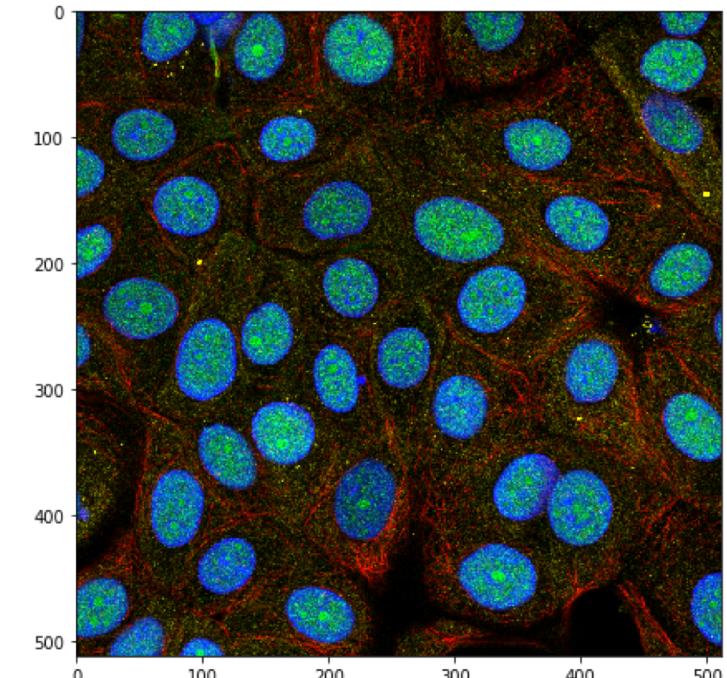
3



Каналы



Фильтры по
каналам

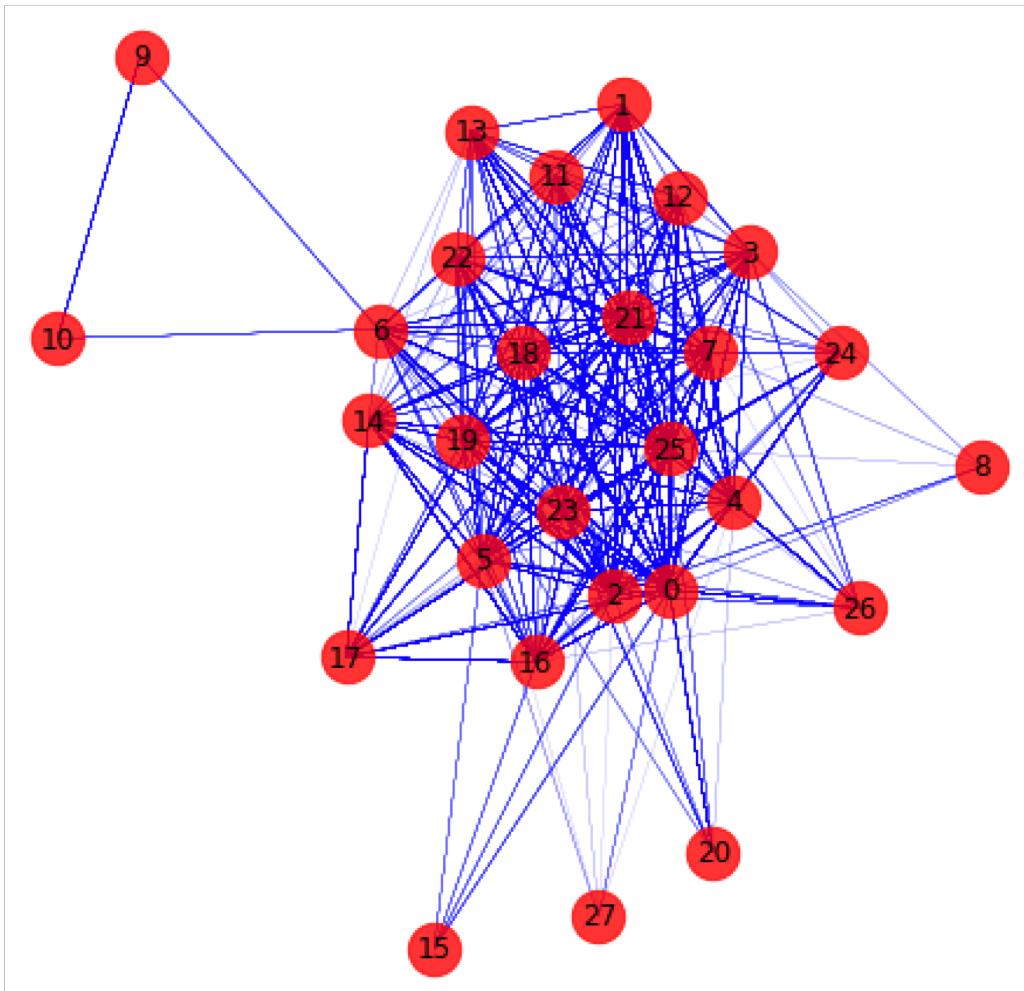


Особенности

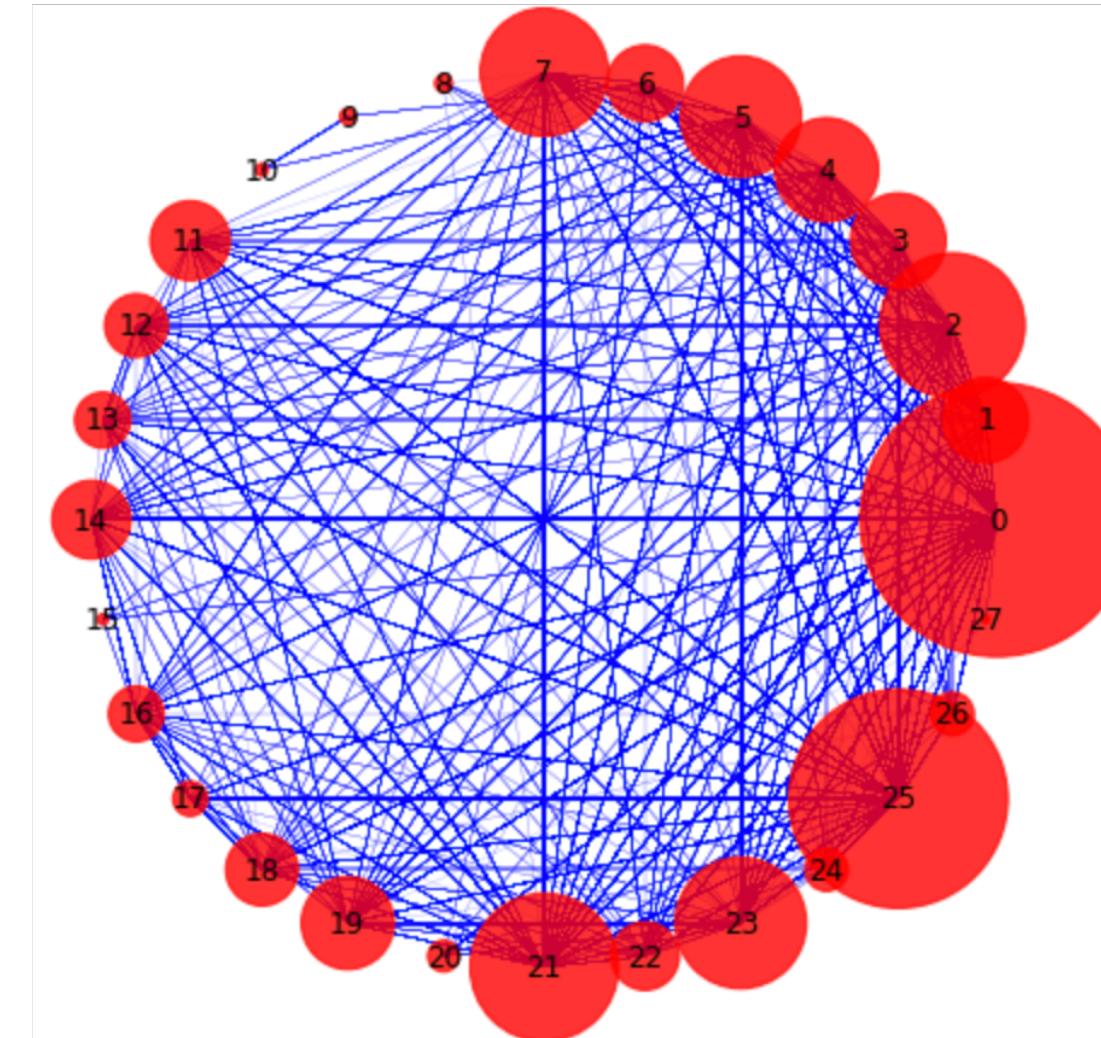
- Некоторые классы представлены очень слабо (пара десятков изображений)
- **Разрешены внешние данные (Мотивация искать лик)**
- 2048x2048 изображения (когда размер GPU памяти важен)
- Некоторые лейблы некорректно размечены (по мнению экспертов)
- Есть явные связи классов

Визуализация связей классов

kamada_kawai_layout

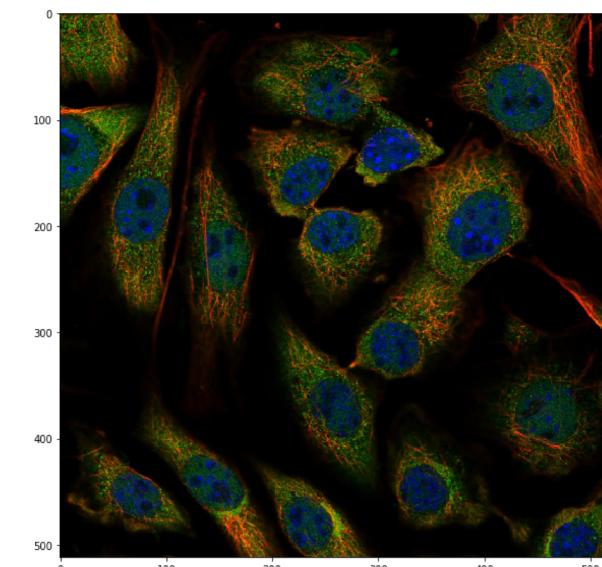
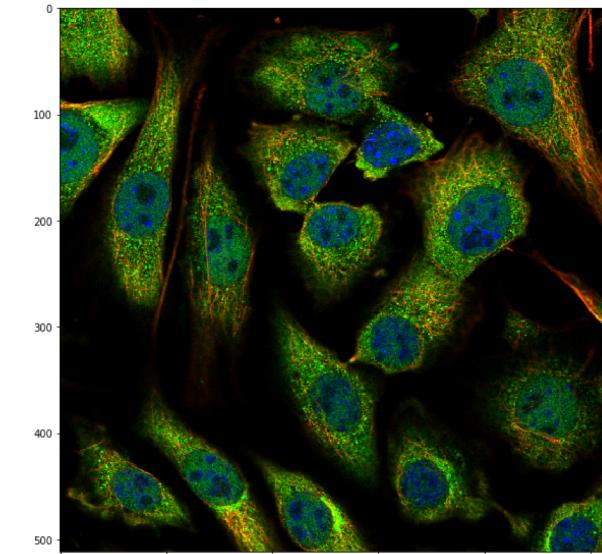


circular_layout



Наш подход

- Скачали внешние данные и нашли дубликаты (на полном наборе данных)
- Валидация (без кросс-валидации)
- Разные архитектуры
- Аугментаций побольше
- Стекинг
- Профит!



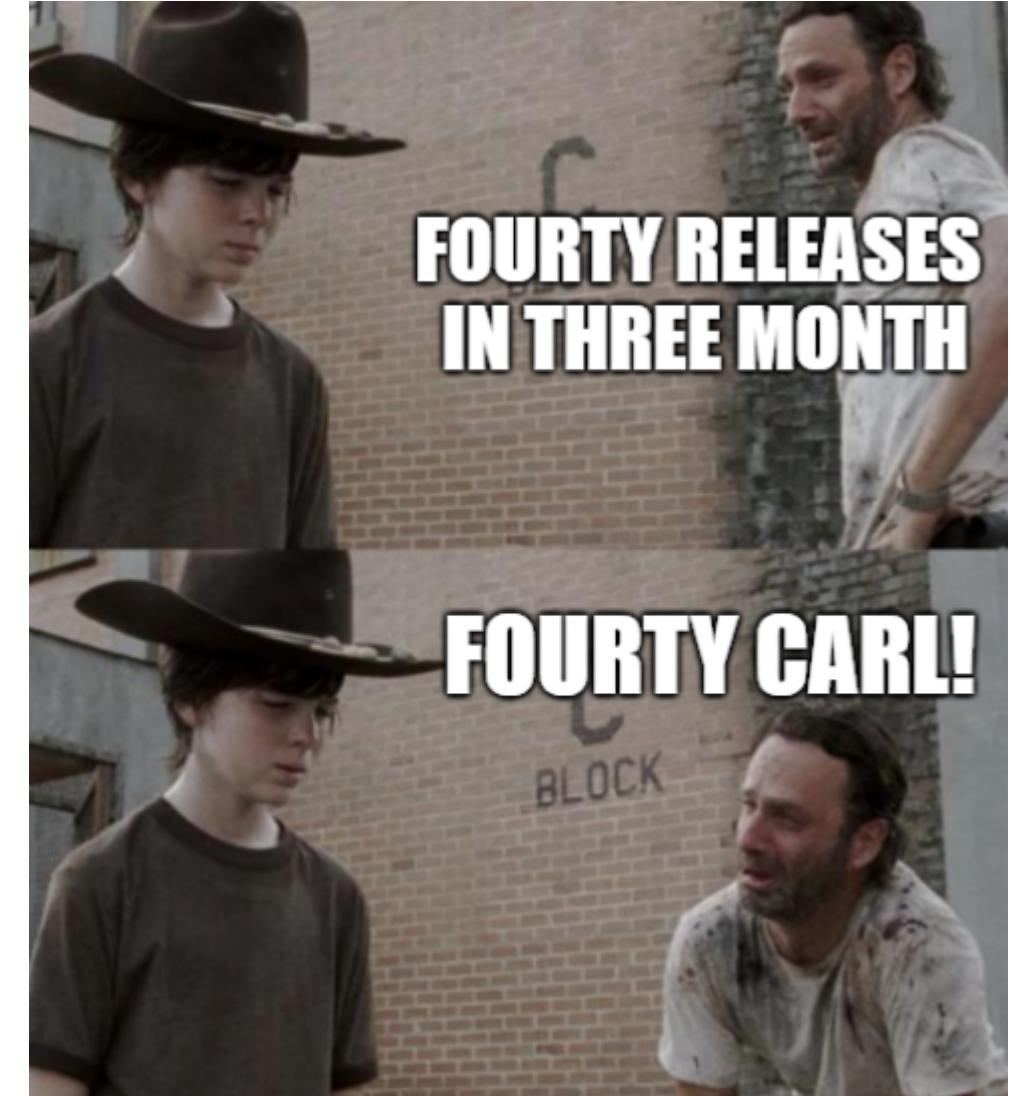
Валидация

- Разделение только на трейн и тест (больше экспериментов с архитектурой и гиперпараметрами)
- **Adversarial validation (Xception)**
- Все “train-train” и “train-external” или в train или в validation
- Все “external-test” дубли в validation
- Для каждого класса в отдельности **5-фолдов (на validation) и усреднение по 20 запускам (стабилизировать трешхолды)**

Fast.ai - «Горшочек, сделай хорошо»

Плюсы

- Все трюки из коробки
 - На базе PyTorch (легко править, заменять архитектуры, дополнять)
-



Минусы

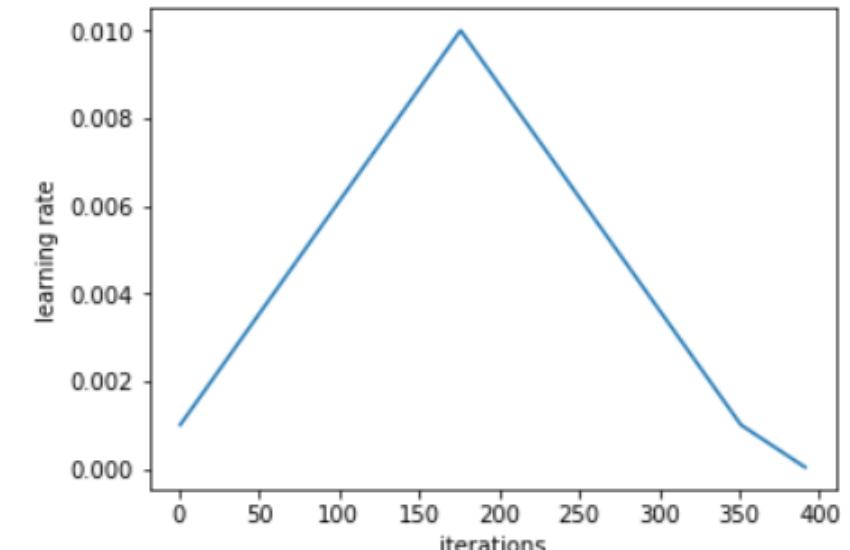
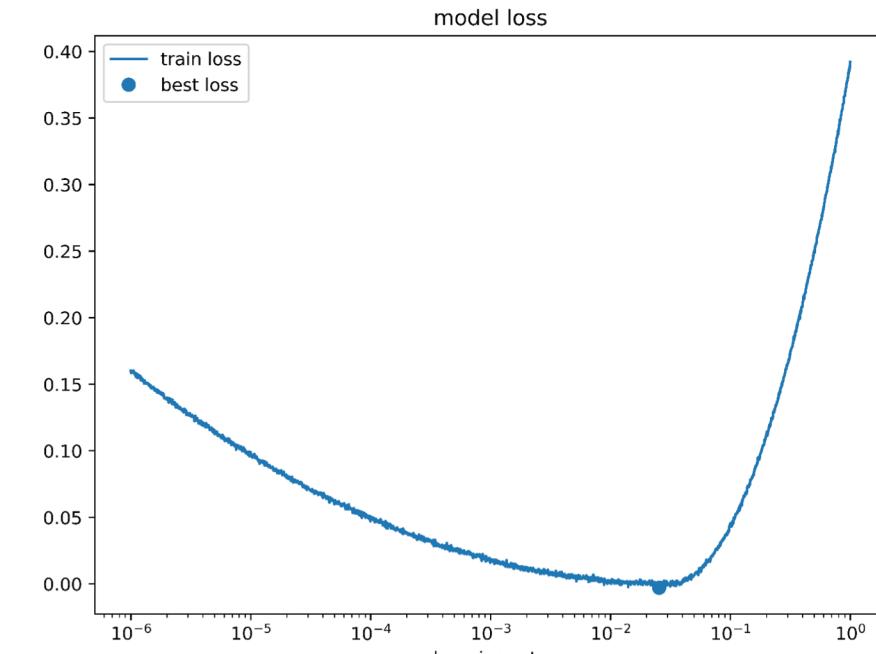
- Специфическое API, местами приходилось патчить (что-то не работало)
- 40 релизов за соревнование!

Сетки

- SE-ResNext-50 – большинство экспериментов (лучший результат)
(InceptionV4, BN-Inception, Xception) – в ансамбле
- Обучение на RGB (RGBY – обучить можно было, но эффекта не было)
- Focal loss, LSEP loss (<https://arxiv.org/abs/1704.03135>)
- Голова сети –
 - AdaptiveConcat
 - BN
 - Dropout
 - Dense
- Аугментации: D4, brightness, warp, crop, resize.
- ТТА: 32x, как и при обучении
- Для стека метафичи: яркость, контрастность, корреляция по каналам

По заветам Джереми

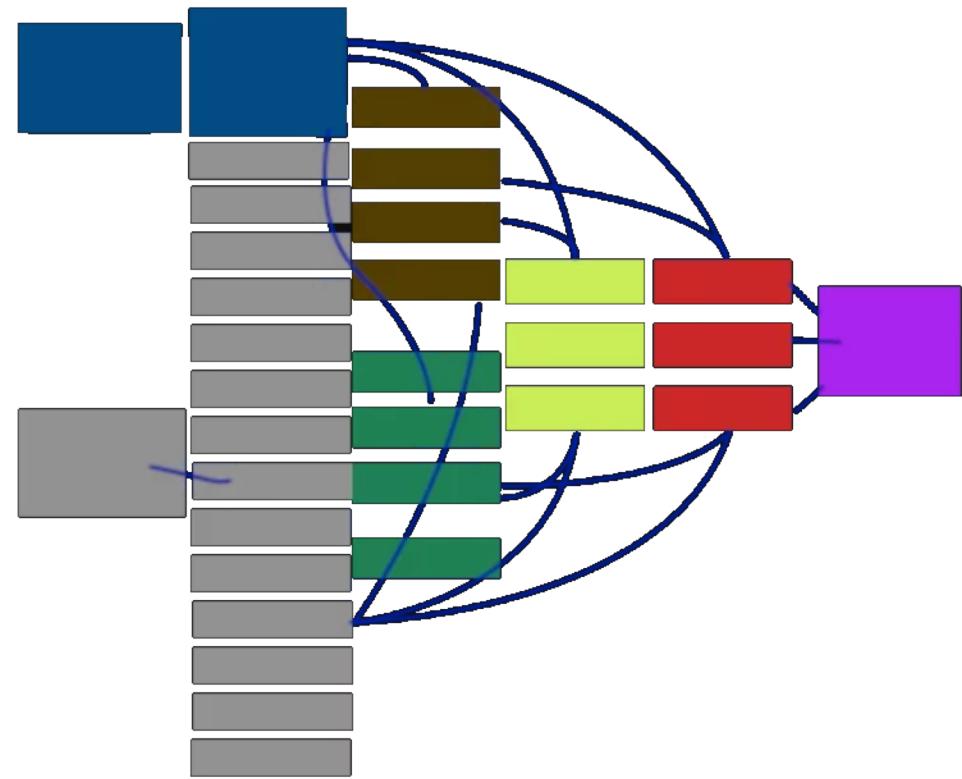
- Подбор lr с использованием
LR finder (<https://arxiv.org/abs/1506.01186>)
- Циклический lr (избегаем локальных минимумов)
- Для эффективного transfer learning-а – затухающий lr по слоям (голова сетки – максимальный – вход минимальный)
- ~~Наркомания (рандом)* lr~~



Стек

- Внешние фолды -5
- Внутренние фолды -3
- 20 раз повторяем
- LightGBM
- Усреднение голосований
- ~375 моделей для каждого класса
- Лучшая модель сама по себе

давала 21 место



Что НЕ сработало (у нас)

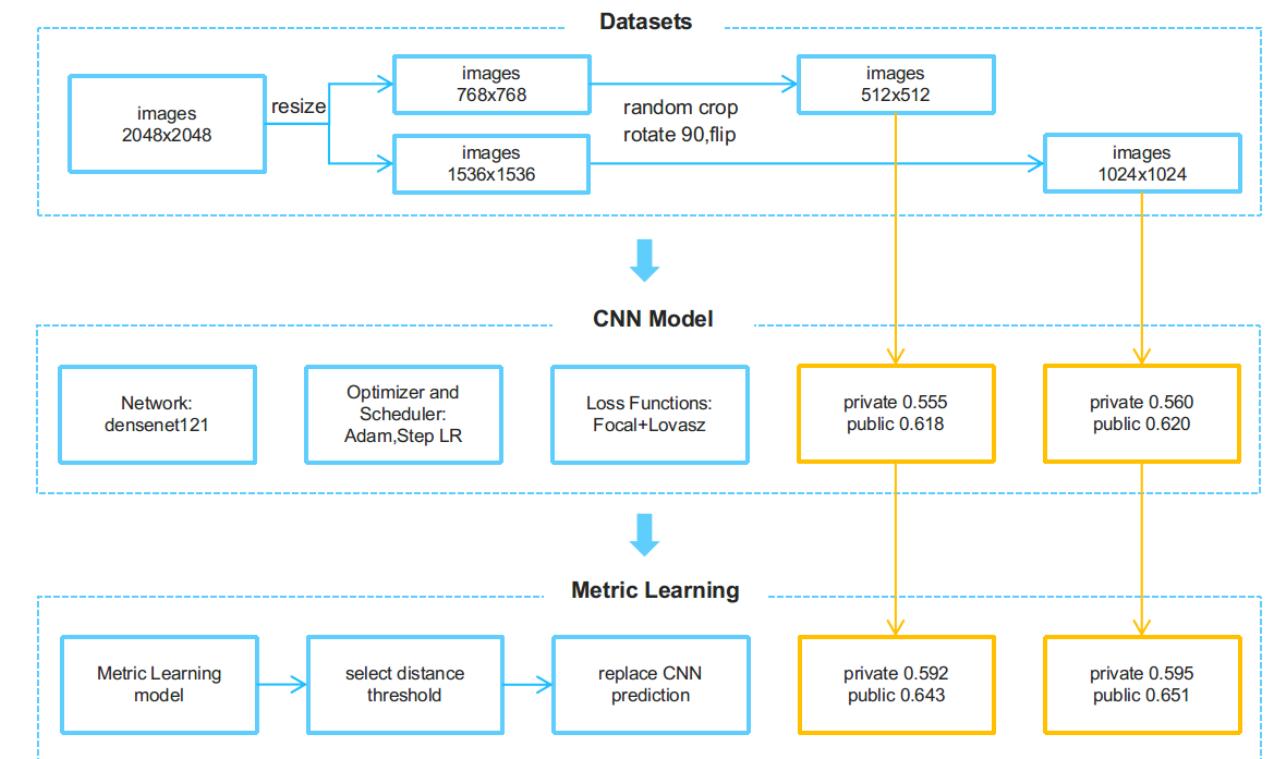
- RGBY (1-у добавили из жалости)
- Sample pairing
- Mixup
- Тренировка one-vs-all моделей
- Аугментации с использованием доменных знаний (зеленого канала)
- Большие сетки (Nasnet, Senet-154) и Малые сетки (resnet18, resnet34)
- Тренировка на 1024
- Snapshot ensembling



А что же bestfitting (1-ое место)

I am sorry for late share, I have worked hard to prepare it in recent days, tried to verify my solution and to make sure it's reproducible, stable, efficient as well as interpretative. (C)

- FocalLoss + **Lovasz**
- Rotate 90, flip, crop 512x512 patches from 768x768 images (or crop 1024x1024 patches from 1536x1536 images)
- densenet121
- Lr scheduling (Adam) [25(30-e5); 30(15e-5); 35(7.5e-5); 40(3e-5); (1e-5)]
- 4x(seed) 512x512 crop (из 768) -> max(prob)
- Metric Learning (+0.03)



Еще интересные моменты (3-е место pudae)

- FocalLoss (gamma=2)
- 1024x1024 (32 gradient accumulation)
- Предикт с усреднением 10 полезных чекпоинтов
- weighted ensemble (3-х моделей)
- ТТА (8)
- Подбор порогов под распределение TP (Lb-probing)

inceptionv3(Public LB 0.574 / Private LB 0.500)

se_resnext50(Public LB 0.583 / Private LB 0.549)

resnet34 (Public LB 0.574 / Private LB 0.500)

Спасибо за внимание!

А также моей Команде и stoически пережившей это соревнование моей жене)