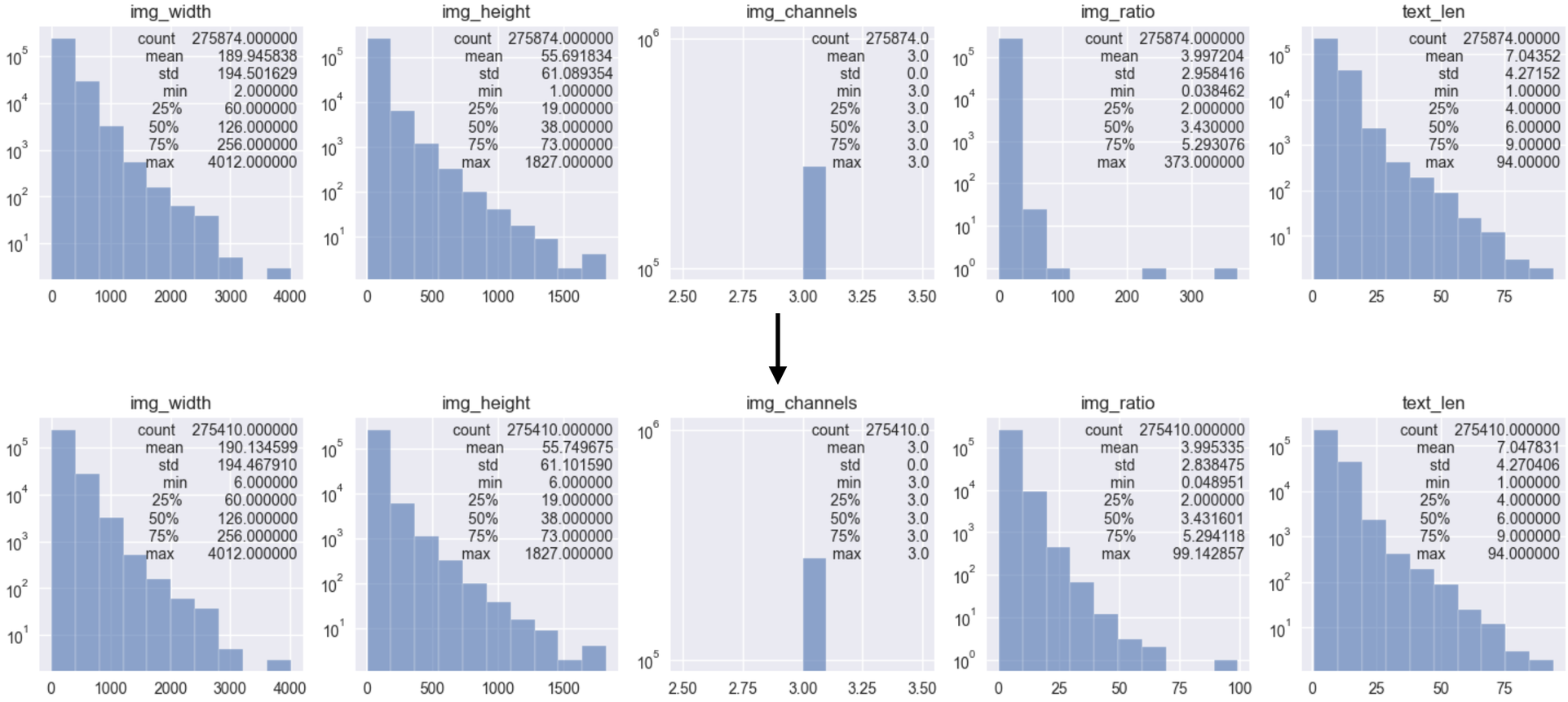
OCR

VK MADE

2023

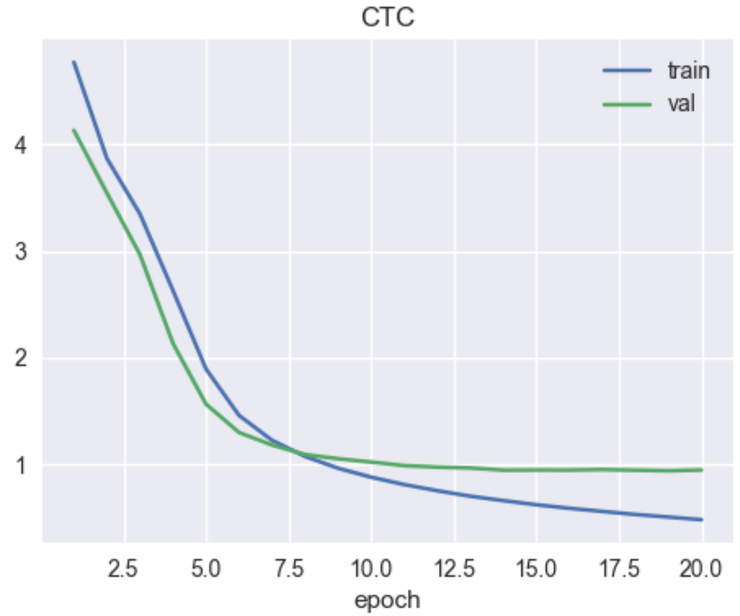
Предобработка данных

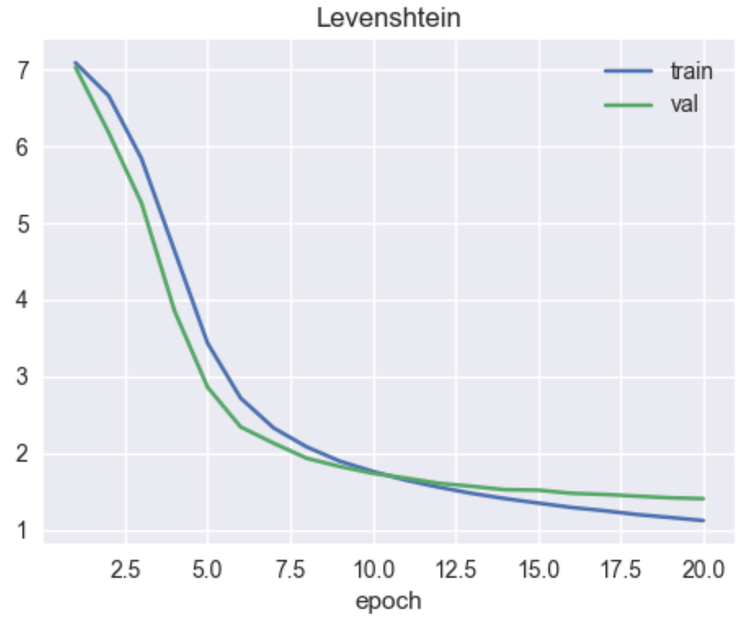
- Из данных были удалены изображения с шириной/высотой менее пяти пикселей.
- Алфавит символов также был уменьшен за счёт удаления изображений с редкими символами, которые встречались менее чем на трёх изображениях.



Модель

- Архитектура молели: CRNN
- Сверточная часть (FeatureExtractor): efficientnet_v2_s (1280 output features) + AvgPool2d (make height equal to 1) + Conv2d (apply projection to increase width to 30)
- Рекуррентная часть (SequencePredictor): GRU (2 layers, bidirectional, hidden size 128)
- Размер входного изображения: 320x64 (все изображения приводятся к этому разрешению)
- Оптимизатор: AdamW
- Число эпох: 20
- Размер батча: 128
- Скорость обучения: 1e-4
- Функция потерь: CTCLoss
- Train/test: 80%/20%
- Аугментация тренировочных данных: Grayscale, Sharpness, GaussianBlur, Rotation





Улучшения результатов

- Первоначально в качестве FeatureExtractor использовалась сеть resnet18, переход на efficientnet_v2_s дал прирост скора на публичном лидерборде с 3.71 до 2.95. Сверточная сеть обучалась вся, ничего не замораживалось.
- В наборе данных было замечено достаточно много изображений с ориентацией текста отличной от привычной, в тесте тоже. Тут прям напрашивался брутфорс, поэтому во время теста каждое изображение поворачивалось дополнительно на 90°, 180° и 270°. Точнее пришлось сделать четыре прогона тестовых данных с разными ориентациями изображений. При каждом прогоне считалась сумма логитов на полученных результатах (степень "уверенности" модели). Как итоговый брался результат с максимальной суммой для одной из четырех ориентаций изображения (включая первичную 0°). Иначе, например, вертикально ориентированный текст тянулся на разрешение 320x64 и сеть пыталась там что-то увидеть. Все это дало итоговое улучшение скора на паблике с 2.95 до 2.45.
- Аналогично была почищена тренировочная разметка, и заново обучена вся модель. В этом случае можно было легко считать расстояние Левенштейна с таргетом и оставлять для обучения ориентацию с минимальным расстоянием. Сильного прироста скора это не дало, 2.41 на паблике🥹
- Можно было ещё прикрутить дополнительно языковую модель и реализовать beam search для формирования предсказаний, это точно улучшило бы результат, но времени не хватило, поэтому во всех экспериментах использовался greedy decoding🫠