

Cyrillic OCR

Разпознаване на ръкописен текст

Задача. Мотивация

- дигитализацията в България (и не само) изостава
 - множество документи все още се попълват на ръка или стари записи не са дигитализирани
 - множество статии и книги, особено по-малко популярни, не са налични в дигитален формат
 - разпознаването на образи - полезно за дипломната ми работа
- създаване на невронна мрежа, която да разпознава символи
 - т.нар. Character Recognition
 - популярен проблем с разнообразни решения
 - все още са възможни подобрения за кирилица
- потенциално надграждане при възможност с цел разпознаване на текст

Задача. Мотивация

- дигитализацията в България (и не само) изостава
 - множество документи все още се попълват на ръка или стари записи не са дигитализирани
 - множество статии и книги, особено по-малко популярни, не са налични в дигитален формат
 - разпознаването на образи - полезно за дипломната ми работа
- създаване на невронна мрежа, която да разпознава ~~символи~~ текст
 - т.нар. ~~Character~~ Text Recognition
 - популярен проблем с разнообразни решения
 - все още са възможни подобрения за кирилица
 - моделът ще разпознава **ръкописен текст**

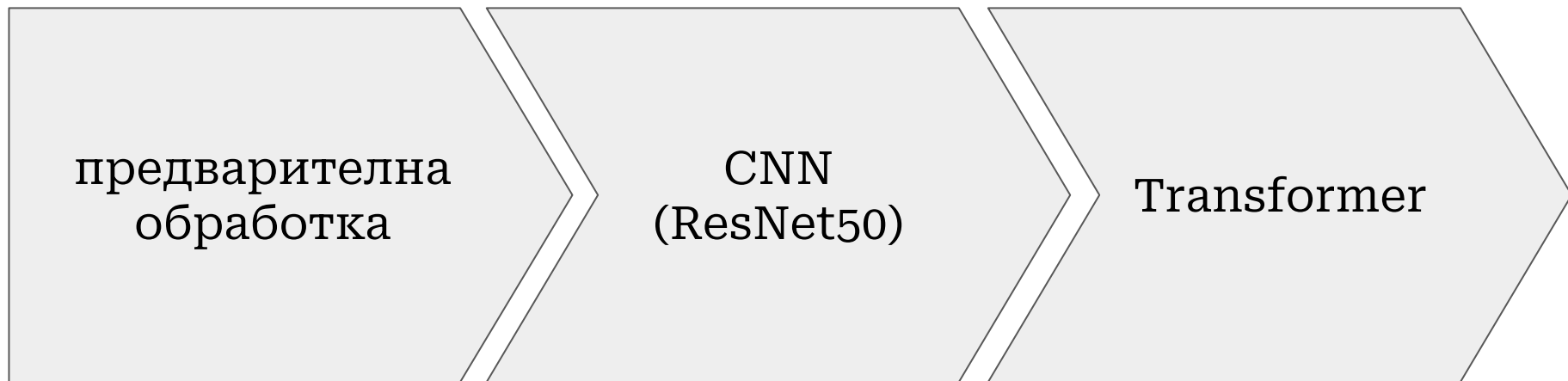
Разпознаване на текст

- Text Detection + Text Recognition + Post-processing
- Text Recognition
 - CNN + RNN + CTC
 - Seq2Seq: CNN енкодер + декодер, който чете отлаво надясно
 - Vision Transformer ViT
 - TrOCR
 - Pix2Seq

Архитектура на Cyrillic OCR

- предварителна обработка
 - трансформация в черно-бяла гама
 - преоразмеряване 256x64
 - превръщане в тензор
 - вкарване на шум в тренировъчния набор - произволно завъртане, контраст, избелване
- CNN
 - ResNet50
 - претренирана
 - имплементация с Deformed CNN - не използвана
- Transformer
 - 6 encoding + 6 decoding layers
 - 8 heads
- крайна обработка
 - превръщане на кодовете в символи

Архитектура на Cyrillic OCR

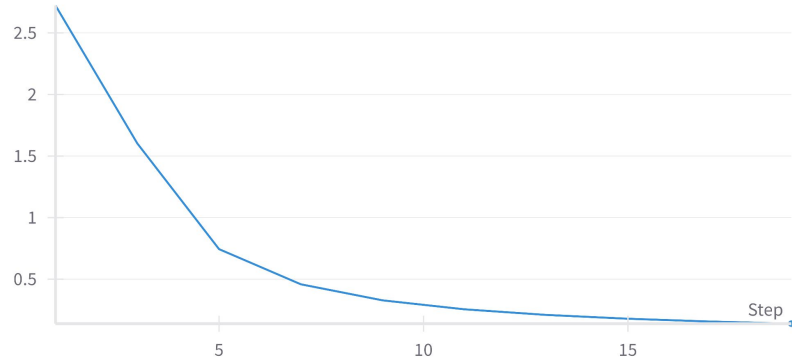


Тренировъчни данни

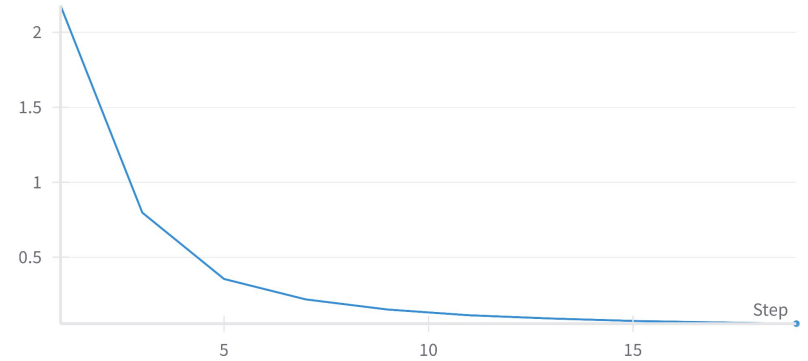
- `pumb-ai/synthetic-cyrillic-large`
 - синтентични данни, генерирани с помощта на криви на Безие
 - 3 800 000 изображения за трениране, валидация и тестване
 - използвани 25% - 702 000 за трениране, 78 000 за валидация
- `constantinwerner/cyrillic-handwriting-dataset`
 - естествени данни от тетрадки и документи
 - 65 152 за трениране, 7 296 за валидация, 1664 за тестване
- съществуват и други, но по-малки и с повече грешки - не са ползвани

Mempuku - pretraining

train_loss

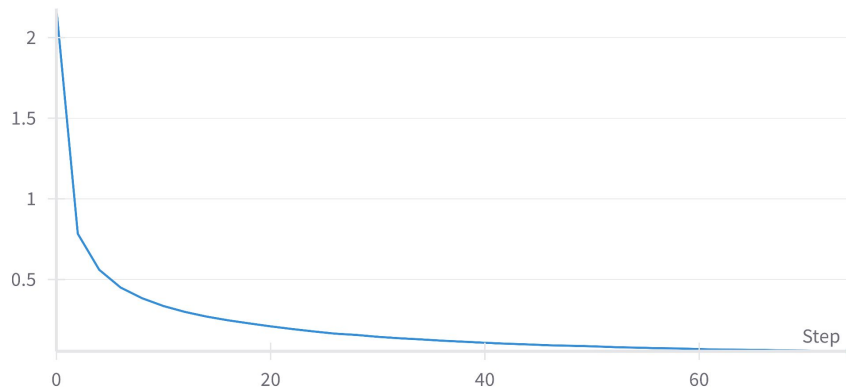


validation_loss

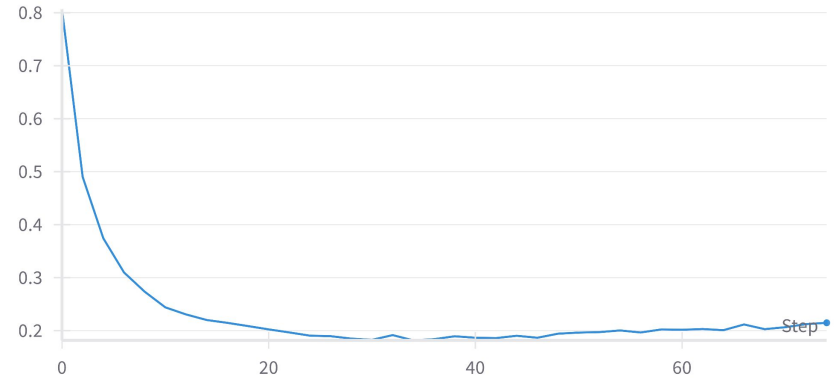


Mempuku - finetuning

train_loss

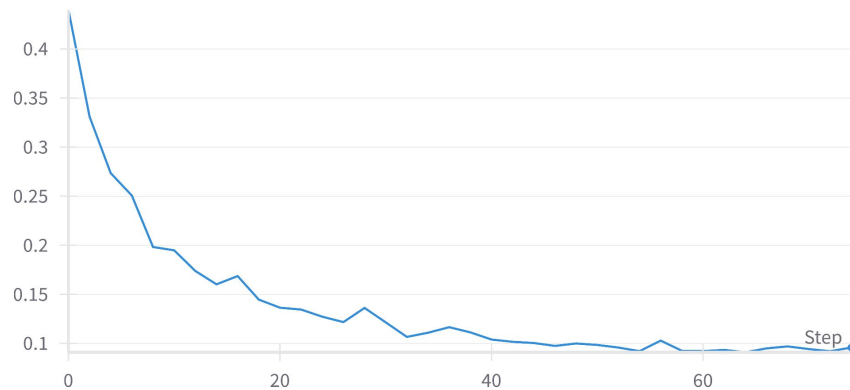


validation_loss

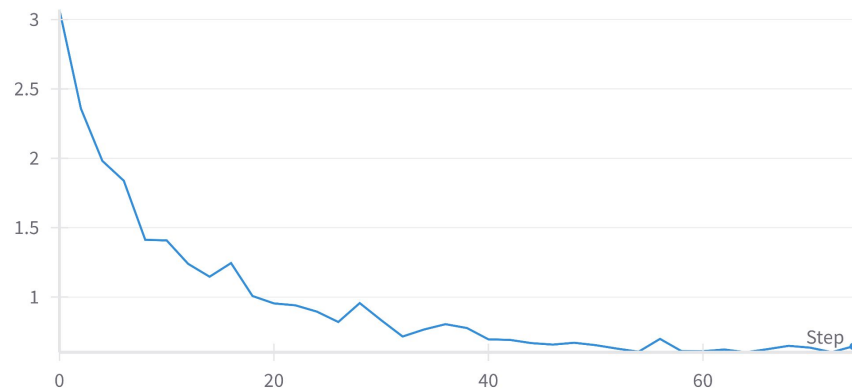


Mempuku - finetuning

cer_loss



wer_loss



Окончателни резултати

32/37

епохи

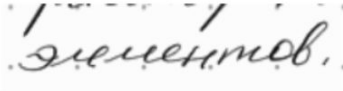
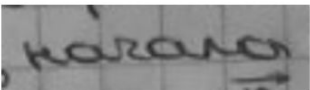
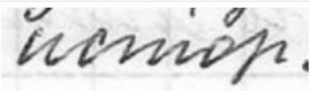
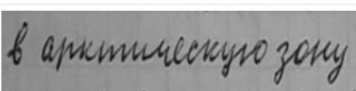
0.15

CER

1.16

WER

Окончательни резултати

Epoch	Image	Ground Truth	Prediction	CER	WER
36		светской	светской	0	0
36		элементов.	элементов.	0	0
37		начала	начала	0	0
37		истор.	истор.	0	0
37		Ласпейреса	Ластей	0.5	5
37		в арктическую зону	в аритическую	0.3333	2

Обобщение

- избраната архитектура демонстрира потенциал за решение на проблема
- ограничени време и средства
- резултатите са успешни с оглед на ограничените ресурси
- съществуват множество възможности за подобрене:
 - експериментиране с различни хиперпараметри
 - 100% използване на синтетичния набор от данни
 - използване на алтернативни набори от данни
 - експериментиране с Deformed Convolution NN
 - комбинация от всички по-горе