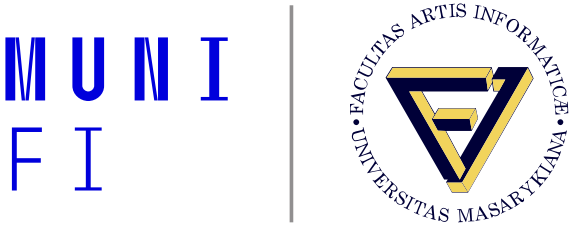


Optical Character Recognition systems for Document Understanding

Mikuláš Bankovič
456421@mail.muni.cz

Faculty of Informatics, Masaryk University

October 10, 2022

Research sources

- A Survey of Deep Learning Approaches for OCR and Document Understanding [5]
- ICDAR 2019 - Scanned Receipts OCR and Information Extraction (SROIE) [2]

Computer Vision problems

Object Detection | Semantic Segmentation | **Instance Segmentation**

Source – Ref. 8

OCR pipeline

Figure 1 given by Subramani et al. [5] shows different approaches to OCR systems.

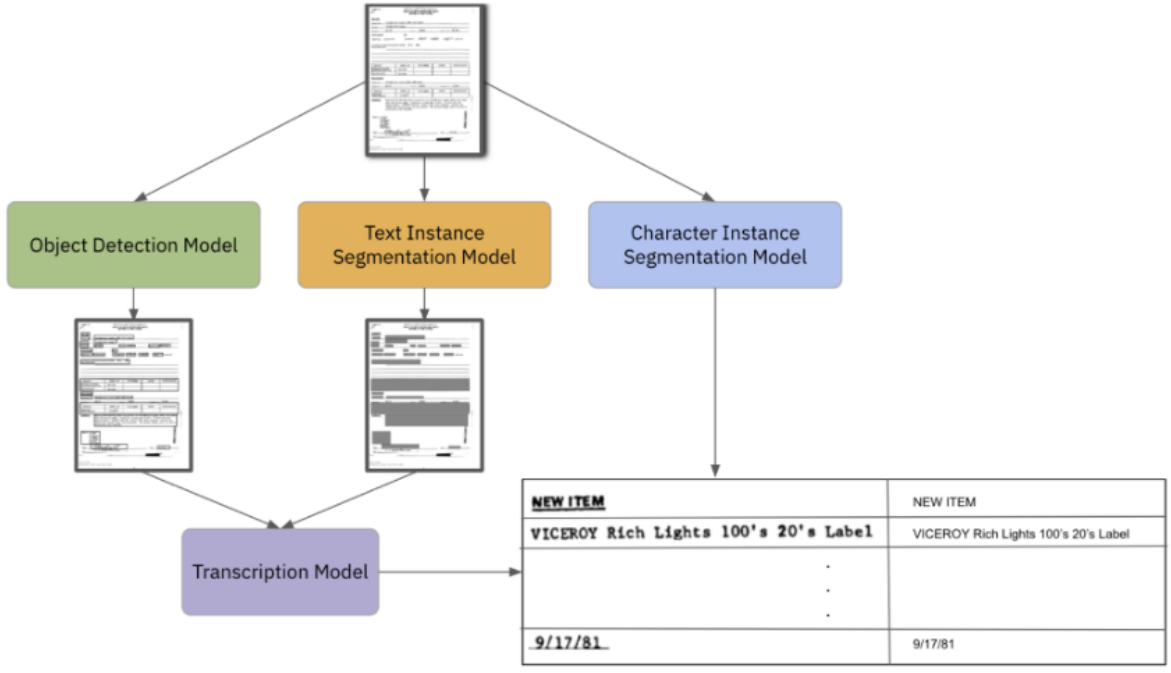


Figure: Object detection and segmentation need transcription(text recognition model)

Text Detection

- CRAFT based on CNN, specifically FCN [1]
- Differentiable Binarization Network (DBNet) [3]

Comparisons

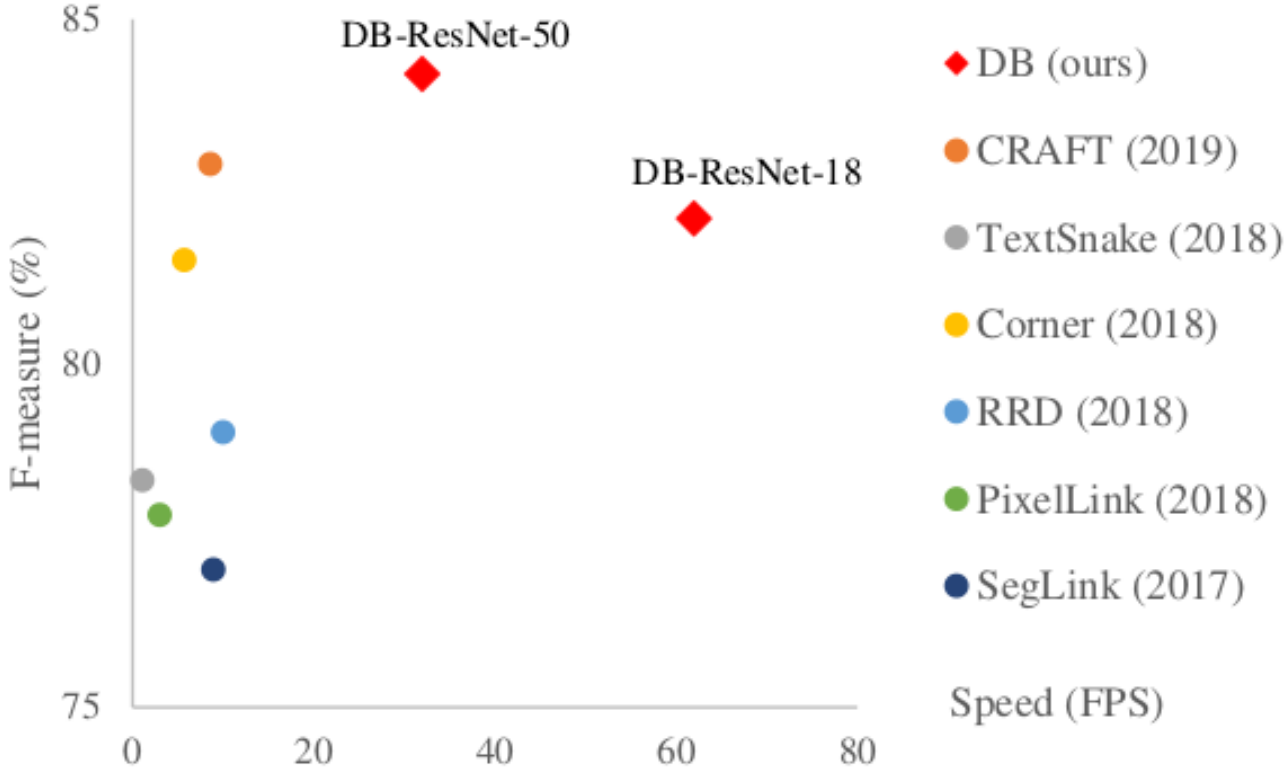


Figure: Speed and metric comparison

Text Recognition

- CRNN with different backbones (mobilenet, vgg)
- Master - CNN + Transformer [4]

Master

Method	Input	Accuracy	Inference Time (ms)	Training Time (h)
SAR [8]	48×160	91.5	16.1	51
MASTER (original)	48×160	95.0	9.2	36
MASTER (improved)	48×160	95.0	4.3	36

Figure: Speed comparison to previous SOTA

Master

- Connectionist Temporal Classification (CTC) loss
- does not require character-level annotations but word-level annotations
- high training parallelization compared to RNN

OCR frameworks

OCR systems allow users to use text detection and text recognition models without having to create own pipeline and visualization tools.

- EasyOCR - many languages - only a few models
- DocTR - newer interface and better variety of models including selected ones

Both have poor documentation and are in their code infancy.

Figure: We do not want to reinvent the wheel

DocTR

- DBNet (pretrained text detection)
- Master (we want to train for our domain)

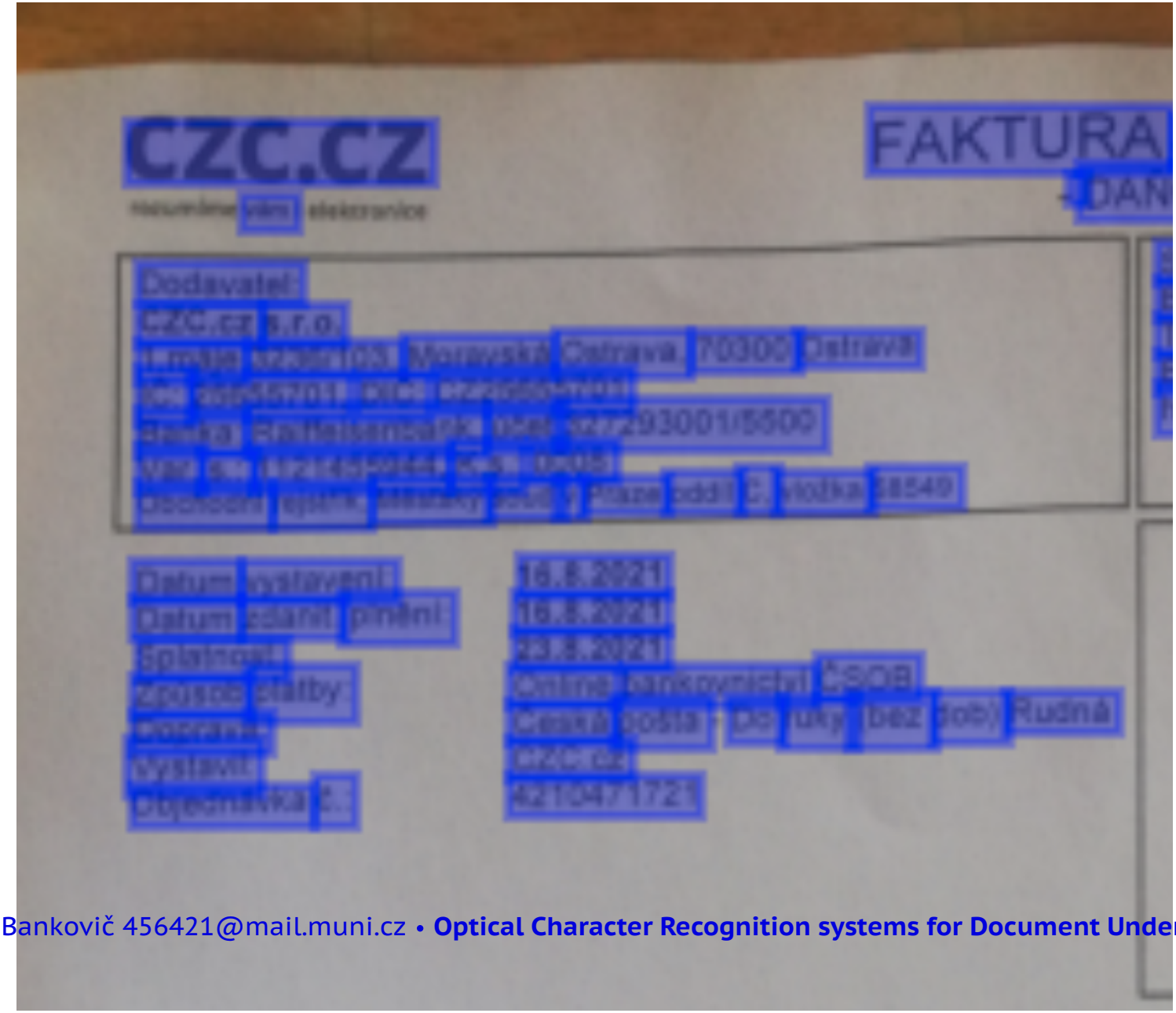
Figure: Doctr

EasyOCR model

```python
import torch.nn as nn
from .modules import ResNet_FeatureExtractor, BidirectionalLSTM

class Model(nn.Module):

    def __init__(self, input_channel, output_channel, hidden_size, num_class):
        super(Model, self).__init__()
        """ FeatureExtraction """
        self.FeatureExtraction = ResNet_FeatureExtractor(input_channel, output_channel)
        self.FeatureExtraction_output = output_channel  # int(imgH/16-1) * 512
        self.AdaptiveAvgPool = nn.AdaptiveAvgPool2d((None, 1))  # Transform final (imgH/16-1) -> 1

        """ Sequence modeling"""
        self.SequenceModeling = nn.Sequential(
            BidirectionalLSTM(self.FeatureExtraction_output, hidden_size, hidden_size),
            BidirectionalLSTM(hidden_size, hidden_size, hidden_size))
        self.SequenceModeling_output = hidden_size

        """ Prediction """
        self.Prediction = nn.Linear(self.SequenceModeling_output, num_class)


    def forward(self, input, text):
        """ Feature extraction stage """
        visual_feature = self.FeatureExtraction(input)
        visual_feature = self.AdaptiveAvgPool(visual_feature.permute(0, 3, 1, 2))  # [b, c, h, w] -> [b, w, c, h]
        visual_feature = visual_feature.squeeze(3)

        """ Sequence modeling stage """
        contextual_feature = self.SequenceModeling(visual_feature)

        """ Prediction stage """
        prediction = self.Prediction(contextual_feature.contiguous())

        return prediction
```

Figure: Source code of so far unnamed model implemented by EasyOCR

Invoice


CZC.CZ
FAKTURA
Dodavatel:
Datum vystavení:
Datum splatnosti:
Způsob platby:
Objednávka č.:
16.8.2021
16.8.2021
23.8.2021
Česká pošta
Rudná

Reconstruction


CZC.CZ

FAKTURA

DA

vàm

Dodavatel:
CZC.cz s.r.o
1.máje 3236/103, Moravská Ostrava 70300 Ostrava
IC 25655701 DIC: CZ25655701
Banka: Raiffeisen ucet 327293001/55
Vars.: 1121455944 K.s 0008
Obchodnrejst Městsk soucV Praze oddiC, vlozk 58549

Datum vystaven 16.8.202
Datum zdani pinèr 16.8.202
Splatnos 23.8.202
Zpusob platb Online bankovnict CSOB
Doprava Ceská post Do ruk (be dob Rudná
Vysta CZC.cz
Objednavk.: 4210471721

Born-Digital Dataset

- Scrape pdf invoices from the internet (provided for us)
- Process pdf files with python libraries (pdfminer, etc) to extract bounding boxes and texts
- Upload on HuggingFace Hub

Born-Digital Dataset

We have 699585 pairs cropped image: extracted text.

Figure: Example of born-digital cropped word. Custom model predicted: 7.1..2018.8.1..2018

Compare OCR on Born Digital dataset

OCR engine	Exact	Partial	No text	FPS
easyocr_generation2	0.88	0.89	0.00	56.13
easyocr_generation1	0.81	0.82	0.01	3.99
doctr_vgg16_bn*	0.78	0.78	0.00	18.93
doctr_mobilenet_v3_large*	0.77	0.78	0.00	40.70
doctr_mobilenet_v3_small*	0.77	0.77	0.00	51.36
tesseract	0.72	0.72	0.24	2.36
doctr_master	0.82	0.82	X	X

Table: Comparison of different recognition networks on born-digital dataset

1

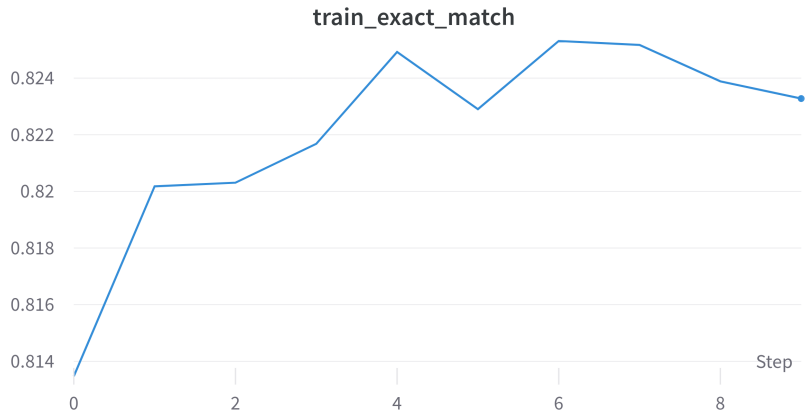
[1] *These are all CRNN architectures with different backbones

Training and validation loss

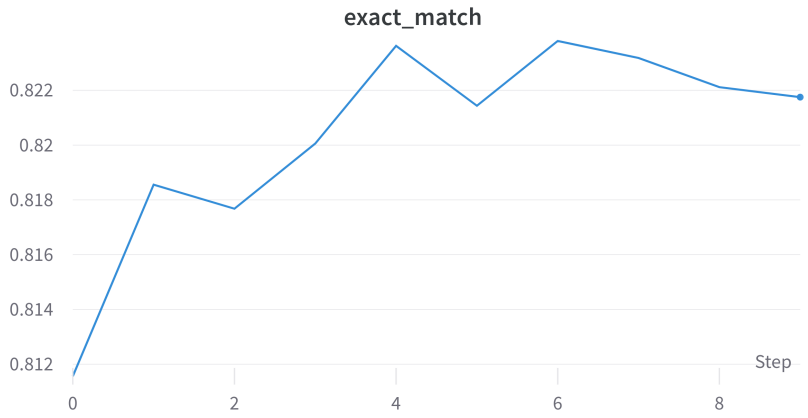
(a) Doctr custom model training loss

(b) Doctr custom model validation loss

Training and validation exact match



(a) Doctr custom model training exact match



(b) Doctr custom model validation exact match

Bibliography I

[1] Youngmin Baek et al. *Character Region Awareness for Text Detection*. 2019. arXiv: 1904.01941 [cs.CV].

[2] *CDAR 2019 Robust Reading Challenge on Scanned Receipts OCR and Information Extraction*. URL: https://rrc.cvc.uab.es/?ch=13&com=introduction (visited on 07/21/2022).

[3] Minghui Liao et al. "Real-time Scene Text Detection with Differentiable Binarization". In: (). URL: http://arxiv.org/abs/1911.08947 (visited on 07/21/2022).

Bibliography II

[4] Ning Lu et al. “MASTER: Multi-aspect non-local network for scene text recognition”. In: (). DOI: 10.1016/j.patcog.2021.107980. URL: https://doi.org/10.1016%5C%2Fj.patcog.2021.107980.

[5] Nishant Subramani et al. “A Survey of Deep Learning Approaches for OCR and Document Understanding”. In: (). URL: https://arxiv.org/abs/2011.13534 (visited on 07/21/2022).

1. We created Czech recognition dataset from invoices
2. We trained prototype custom model architecture
3. We compared pretrained solution and evaluated our prototype

MUNI
FACULTY
OF INFORMATICS