OpenPSS

An Open Page Stream Segmentation Benchmark

Ruben van Heusden, Jaap Kamps & Maarten Marx

25-09-2024

Information Retrieval Lab, University of Amsterdam

UNIVERSITY OF AMSTERDAM
Informatics Institute

IR
Lab
AMS

Introduction

What is Page Stream Segmentation (PSS)?

Origins

- Document digitization through **scanning**
- Combining multiple documents into 1 PDF file (**cost-effective**)

Problem

- Issues with downstream tasks (**searching**)

Task

- Automatically recovering document boundaries
- Usage of text- and/or images (**multimodal**)

Challenges & Motivation

Challenges

- Almost no (large) public datasets
- Different evaluation metrics in different papers :**How to compare?**

Goals

- Providing an overview of current PSS methods
  - Two large multimodel datasets
  - Wider set of evaluation metrics for model evaluation

The Benchmark

The benchmark

Datasets

- 2 large multimodal Dutch datasets obtained from FOIA requests totalling roughly **400 streams**, **32,000 documents** and **140,000 pages** (**SHORT** and **LONG**)

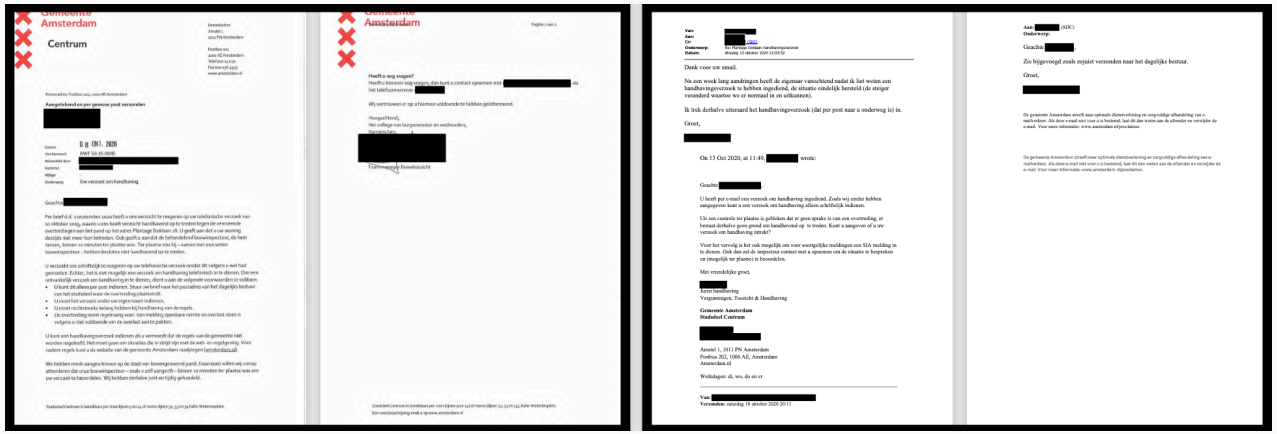


Figure 1: Example of a stream of two documents, both with two pages

Evaluation Metrics

Evaluation

- Apart from Precision, Recall and F1, using **Panoptic Quality** [2]
- Allows for model to receive credit for partially correct predictions

Experiments

What models did we use?

Model Types

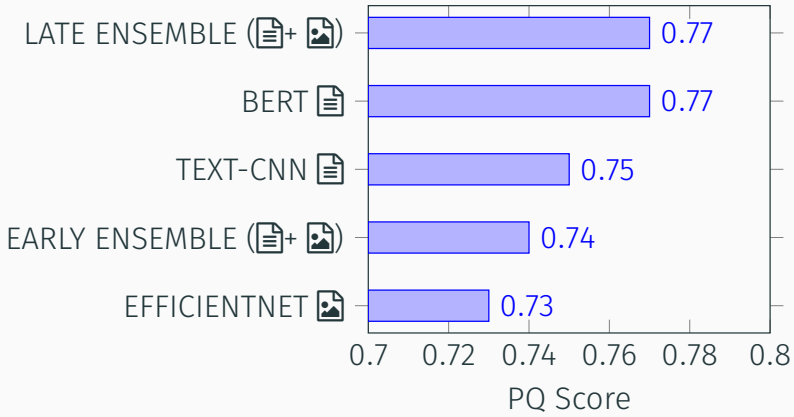
- Simple baselines (KNN, XGBoost) with pre-trained embeddings
- Neural text- and/or vision models (VGG16, BERT)[4, 1]
- Sequence-tagging models (LSTM) [3]

Task Variants

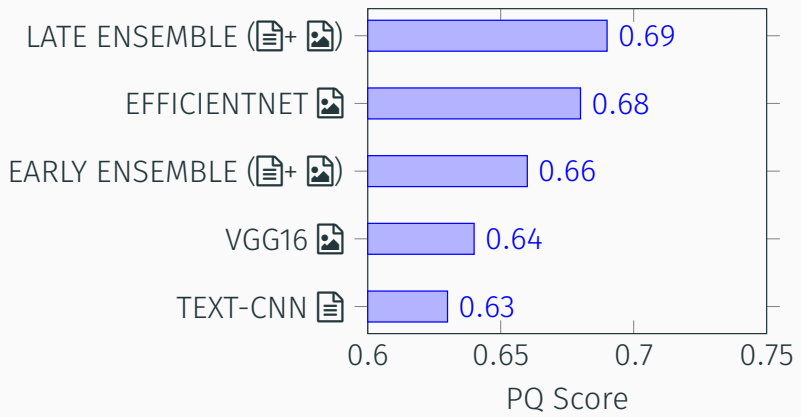
- **Classic** Train and test on the same dataset
- **Robust** Train on dataset A, test on dataset B (**more realistic**)
  - Measure average performance drop

Result Highlights

Standard Setting



(a) LONG Dataset



(b) SHORT Dataset

- For both datasets (multimodal) neural methods are superior
- Text methods better on **LONG** dataset, image methods better on **SHORT**

Robust Setting

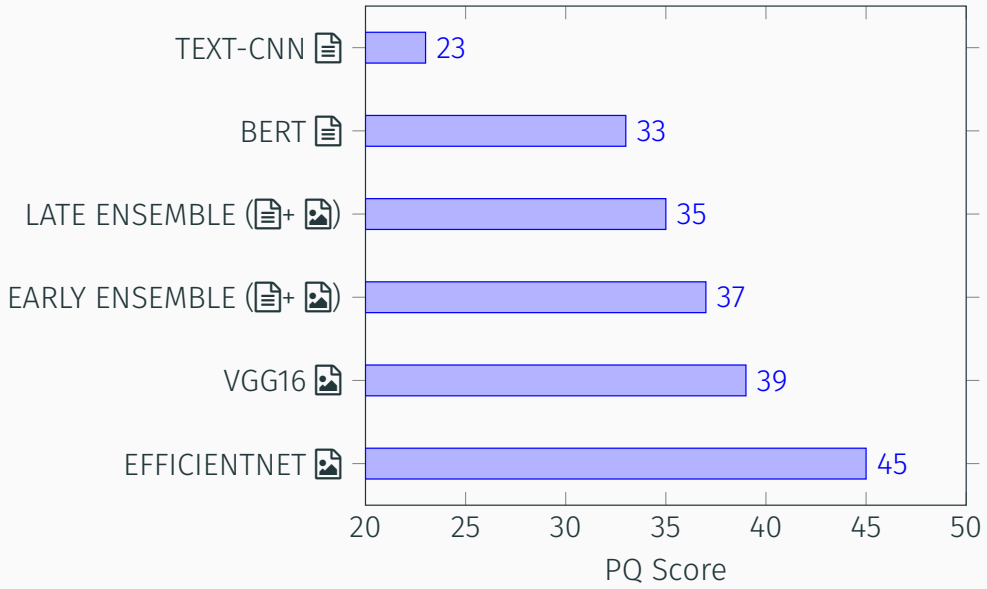


Figure 3: Average performance drop in percentage

- Text models are (generally) the most robust
- Image models are not as robust as text or ensemble

Practical Implications

Types of Mistakes

- 8 out of 10 hits is a TP, on average less than 2 pages off
- **Remaining hits**
  - Contain multiple documents
  - Contain only part of document

Conclusion

Conclusion & Future Work

Conclusions

- We constructed the **OpenPSS** benchmark
- Extensively compared 17 baselines on both datasets
  - (Multimodal) neural methods superior
  - neural text models most robust
- Analysis of impact on a search engine

Future Work

- Expanding on sequence labelling methods (**Transformers**)

Questions?

References i


F. A. Braz, N. C. da Silva, and J. A. S. Lima.
Leveraging effectiveness and efficiency in page stream deep segmentation.
Eng. Appl. of Artif. Intell., 105:104394, 2021.


A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár.
Panoptic segmentation.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CCVPR)*, pages 9404–9413, 2019.


O. Koshorek, A. Cohen, N. Mor, M. Rotman, and J. Berant.
Text segmentation as a supervised Learning task.
In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 469–473,

New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

G. Wiedemann and G. Heyer.
Multi-modal page stream segmentation with convolutional neural networks.
Lang. Resour. and Evaluation, 55(1):127–150, 2021.

Get in touch!

- email: r.j.vanheusden@uva.nl
- website: `https://staff.fnwi.uva.nl/r.j.vanheusden`

UNIVERSITY OF AMSTERDAM
Informatics Institute

IR Lab AMS