International Journal on Document Analysis and Recognition (IJDAR) (2023) 26:241–254
https://doi.org/10.1007/s10032-023-00437-8

SPECIAL ISSUE PAPER

Historical document image analysis using controlled data for pre-training

Najoua Rahal[1] · Lars Vögtlin[1] · Rolf Ingold[1]

Received: 14 November 2022 / Revised: 15 March 2023 / Accepted: 16 April 2023 / Published online: 10 May 2023

Abstract
Using neural networks for semantic labeling has become a dominant technique for layout analysis of historical document images. However, to train or fine-tune appropriate models, large labeled datasets are needed. This paper addresses the case when only limited labeled data are available and promotes a novel approach using so-called controlled data to pre-train the networks. Two different strategies are proposed: The first addresses the real labeling task by using artificial data; the second uses real data to pre-train the networks with a pretext task. To assess these strategies, a large set of experiments has been carried out on a text line detection and classification task using different variants of U-Net. The observations, obtained from two different datasets, show that globally the approach reduces the training time while offering similar or better performance. Furthermore, the effect is bigger on lightweight network architectures.

Keywords Historical document images · Layout analysis · Neural networks · Training strategies

1 Introduction

In the last decade, deep neural networks have become the dominant technology for image analysis and recognition. Enormous progress has been reported for various tasks such as image classification [1], segmentation [2], and object detection [3]. This observation also applies to historical document image analysis, a research field that is gaining increasing importance, given the tremendous demand of historians and scholars to analyze large sets of digitized documents [4].

For document image analysis applications, layout analysis plays a central role. One of the most important tasks of layout analysis is text line and text block detection, as it is a useful preliminary step for automatic text recognition. Additionally, various scripts and text styles can occur in the same document, so it is interesting to classify text lines correctly. Thus, this paper focuses on so-called *text line detection and classification* in historical documents. An efficient way to address this goal is to consider it a semantic labeling task, which can be solved with deep neural networks of the U-Net family.

The text line detection task has been addressed by many researchers, and respectable results are regularly reported for experiments made with data sets produced by the research community for which enough ground-truth information is available. However, in practice, the models trained with these datasets are hardly adaptable to process other document classes for which only limited ground-truth annotations are available.

To overcome the above difficulties, we propose a novel approach based on transfer learning, using *controlled data* for pre-training. Two complementary strategies are investigated: the first uses artificial data and pre-train networks to solve the real task. The second strategy involves pre-training the networks with a pretext task applied to real data using self-supervised training.

The artificial data we use are synthetically generated, including the necessary annotations; they can be automatically produced in large quantities. Instead of producing synthetic data that imitates the real data, e.g., using generative adversarial networks [5], we create them in an entirely

N. Rahal, L. Vögtlin and R. Ingold contributed equally to this work.

✉ Najoua Rahal
najoua.rahal@unifr.ch

Lars Vögtlin
lars.voegtlin@unifr.ch

Rolf Ingold
rolf.ingold@unifr.ch

[1] Document Image and Voice Analysis Group (DIVA), University of Fribourg, 1700 Fribourg, Switzerland