International Journal on Document Analysis and Recognition (IJDAR) (2023) 26:187–209
https://doi.org/10.1007/s10032-023-00429-8

SPECIAL ISSUE PAPER

Analyzing the potential of active learning for document image classification

Saifullah Saifullah[1,2] · Stefan Agne[1,3] · Andreas Dengel[1,2] · Sheraz Ahmed[1,3]

Received: 14 November 2022 / Accepted: 23 March 2023 / Published online: 25 April 2023

Abstract

Deep learning has been extensively researched in the field of document analysis and has shown excellent performance across a wide range of document-related tasks. As a result, a great deal of emphasis is now being placed on its practical deployment and integration into modern industrial document processing pipelines. It is well known, however, that deep learning models are data-hungry and often require huge volumes of annotated data in order to achieve competitive performances. And since data annotation is a costly and labor-intensive process, it remains one of the major hurdles to their practical deployment. This study investigates the possibility of using active learning to reduce the costs of data annotation in the context of document image classification, which is one of the core components of modern document processing pipelines. The results of this study demonstrate that by utilizing active learning (AL), deep document classification models can achieve competitive performances to the models trained on fully annotated datasets and, in some cases, even surpass them by annotating only 15–40% of the total training dataset. Furthermore, this study demonstrates that modern AL strategies significantly outperform random querying, and in many cases achieve comparable performance to the models trained on fully annotated datasets even in the presence of practical deployment issues such as data imbalance, and annotation noise, and thus, offer tremendous benefits in real-world deployment of deep document classification models. The code to reproduce our experiments is publicly available at https://github.com/saifullah3396/doc_al.

Keywords Document image classification · Document analysis · Active learning · Deep active learning

1 Introduction

Document analysis is a field of research that deals with automating the process of reading, analyzing, and understanding business documents. Modern businesses rely heavily on business documents to communicate details of their internal and external transactions, which is critical to their efficiency and productivity. As large volumes of documents are produced on a daily basis, there is an urgent need today to automate the processing of these documents to facilitate tasks such as search, retrieval, and information extraction. However, automated processing of documents can be particularly challenging for a number of reasons, including high levels of data complexity [1], large inter-class similarity and intra-class variance [2], and corruption of scanned document data with various types of distortions [3].

To address the aforementioned challenges, deep learning has been extensively explored in the field and has proven to be exceptionally effective in a wide range of document analysis tasks such as document image classification [4, 5], layout analysis [5], OCR [6], etc. However, deep learning presents some unique challenges of its own. One major disadvantage of deep learning-based approaches is that their performance is heavily dependent on the availability of large amounts of annotated training data. While most real-world tasks have a vast amount of data available that could

✉ Saifullah Saifullah
saifullah.saifullah@dfki.de

Stefan Agne
stefan.agne@dfki.de

Andreas Dengel
andreas.dengel@dfki.de

Sheraz Ahmed
sheraz.ahmed@dfki.de

[1] German Research Center for Artificial Intelligence, 67663 Kaiserslautern, Germany

[2] RPTU Kaiserslautern-Landau, 67663 Kaiserslautern, Germany

[3] DeepReader GmbH, 67663 Kaiserslautern, Germany