Nougat: Neural Optical Understanding for Academic Documents

Lukas Blecher* **Guillem Cucurull** **Thomas Scialom** **Robert Stojnic**

Meta AI

arXiv:2308.13418v1 [cs.LG] 25 Aug 2023

Abstract

Scientific knowledge is predominantly stored in books and scientific journals, often in the form of PDFs. However, the PDF format leads to a loss of semantic information, particularly for mathematical expressions. We propose Nougat (**N**eural **O**ptical **U**nderstandin**g** for **A**cademic Documen**t**s), a Visual Transformer model that performs an *Optical Character Recognition* (OCR) task for processing scientific documents into a markup language, and demonstrate the effectiveness of our model on a new dataset of scientific documents. The proposed approach offers a promising solution to enhance the accessibility of scientific knowledge in the digital age, by bridging the gap between human-readable documents and machine-readable text. We release the models and code to accelerate future work on scientific text recognition.

1 Introduction

The majority of scientific knowledge is stored in books or published in scientific journals, most commonly in the Portable Document Format (PDF). Next to HTML, PDFs are the second most prominent data format on the internet, making up 2.4% of common crawl [1]. However, the information stored in these files is very difficult to extract into any other formats. This is especially true for highly specialized documents, such as scientific research papers, where the semantic information of mathematical expressions is lost.
Existing Optical Character Recognition (OCR) engines, such as Tesseract OCR [2], excel at detecting and classifying individual characters and words in an image, but fail to understand the relationship between them due to their line-by-line approach. This means that they treat superscripts and subscripts in the same way as the surrounding text, which is a significant drawback for mathematical expressions. In mathematical notations like fractions, exponents, and matrices, relative positions of characters are crucial.
Converting academic research papers into machine-readable text also enables accessibility and searchability of science as a whole. The information of millions of academic papers can not be fully accessed because they are locked behind an unreadable format. Existing corpora, such as the S2ORC dataset [3], capture the text of 12M[2] papers using GROBID [4], but are missing meaningful representations of the mathematical equations.

To this end, we introduce Nougat, a transformer based model that can convert images of document pages to formatted markup text.
The primary contributions in this paper are

- Release of a pre-trained model capable of converting a PDF to a lightweight markup language. We release the code and the model on GitHub[3]
- We introduce a pipeline to create dataset for pairing PDFs to source code
- Our method is only dependent on the image of a page, allowing access to scanned papers and books

*Correspondence to: lblecher@meta.com
[2]The paper reports 8.1M papers but the authors recently updated the numbers on the GitHub page https://github.com/allenai/s2orc
[3]https://github.com/facebookresearch/nougat