Internship proposal: "Converting math formula images to LaTeX encoding with machine learning"

Chair of research and teaching in artificial intelligence

Keywords: *Optimization, calibration, cross-validation, sparse regression, MCP*

Context

LaTeX[1] (stylised as LaTeX) is a typesetting system and document edition framework based on a markup language of the same name [2]. Descriptive markup languages as LaTeX (or Markdown, HTML) allows to decouple the structure and content of the document from its rendering, thanks to tags and structure markers. Document writing and edition is done by describing the material conceptually, rather than visually as in standard "What You See Is What You Get" (WYSIWYG) document edition software[2].

LaTeX is widely used in the scientific community to write and edit scientific documents, including articles, manuscripts, presentation slides, etc., in particular because it allows to easily write and render mathematical formulae. Here is an example of how the Navier-Stokes equation is written in LaTeX:

```
\begin{align*}
    \frac{\partial \mathbf{u}}{\partial t}
    + (\mathbf{u} \cdot \nabla) \mathbf{u}
    =
    - \frac{1}{\rho}\nabla P + \nu \nabla^2 \mathbf{u}
\end{align*}
```

which is rendered as:

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla)\mathbf{u} = -\frac{1}{\rho}\nabla P + \nu \nabla^2 \mathbf{u}$$

When writing scientific materials and especially mathematical materials, one often needs to present or use previous results and researches, including mathematical formulae presented in existing articles or manuscripts. Copying an existing (and sometimes complex) mathematical formula and converting it into the corresponding LaTeX encoding can be a cumbersome task. In this context, an *automatic math formula images to LaTeX encoding converter* would be a very useful tool to *improve writing efficiency* and *avoid copy error*. Working with images of mathematical formulae originating from hand-written documents or digital documents (e.g. manuscript PDFs, web pages, etc.), such a tool could automatically generate the corresponding LaTeX code.

Based on *Optical Character Recognition* (OCR) approaches (see [6], [7] for a review on this more general subject), multiple software solutions exist to complete this purpose. Many commercial (and proprietary)

[1] https://www.latex-project.org/

[2] *e.g.,* standard (proprietary or free/open-source) office suites

online web applications are now available like the famous Mathpix[3] or ScribbleMyScience[4], as well as proprietary software like InftyReader[5] (2021), or free and open source software such as the older Freehand Formula Entry System[6] (FFES, 2007, GPL v2 license, see [3] for a review).

Converting math formula images to LaTeX encoding is an active research subject, especially regarding machine learning based methods, following the development of deep learning for character recognition in the recent years. Many contributions to the subject can be found including:

- publications likes [1] with various corresponding implementations[7,8,9], [9] [10],
- many proof-of-concept projects, see a non-exhaustive list of github repository on the subject[10] or some student projects ([4][11], [5]),
- more mature sofware with screenshot capacity like LaTeX-OCR[12].

Supervised machine learning methods are based on a model training step that requires collections of labeled data, *i.e.,* (in our context) data providing examples of math formula images and related formula LaTeX encoding. The larger the training datasets are, the better the conversion accuracy for new images will be. Luckily, numerous sources for LaTeX codes containing math formulae can be found on the Internet, *e.g.,* in the arXiv[13] prepublication database, on Wikipedia[14], or in the imag2latex-100k dataset[15]. To robustify images conversion, especially to improve the conversion of poor quality ones, data augmentation is often leveraged (*e.g.,* [8]).

Objectives and internship conditions

The objectives of this internship are manifold and include:

- a review of the literature and existing open source solutions for math formula image to LaTeX encoding conversion,
- improving an existing solution or developing a new method from scratch (depending on the performance and implementation of existing solutions) to efficiently convert math formula image to LaTeX encoding,
- kick-start a dynamic (and a community-based collaborative project) to offer a reliable and efficient alternative to commercial solution for the scientific community.

During this internship, you will strengthen your knowledge and expertise about deep learning for OCR, deep learning method implementation, but also software development, including programming, software project management, collaborative development, and more. You will also discover the functioning of an academic research laboratory. Eventually, you will contribute to build and/or improve a free and open-source software tools, which will be very useful for the scientific community.

The intern is expected to have basic knowledge in machine learning and deep learning (potentially also in OCR but it is not mandatory), basic knowledge and interest for programming (Python at least) and

[3] https://mathpix.com/
[4] https://scribblemyscience.com/
[5] https://www.sciaccess.net/en/InftyReader/
[6] https://www.cs.rit.edu/~rlaz/ffes/
[7] https://github.com/harvardnlp/im2markup/ (original code, written in Lua + Python + Torch, MIT license)
[8] https://github.com/ritheshkumar95/im2latex-tensorflow (TensorFlow reimplementation, no license)
[9] https://github.com/yixuanzhou/image2latex (PyQt5 GUI, no license)
[10] https://github.com/topics/im2latex
[11] https://github.com/guillaumegenthial/im2latex
[12] https://github.com/lukas-blecher/LaTeX-OCR (2021, MIT license, CLI or GUI)
[13] https://arxiv.org/
[14] https://www.wikipedia.org/
[15] https://zenodo.org/record/56198#.V2px0jXT6eA

software development, but also an interest for the academic research and to contribute to open science.

The internship will take place at the *Institut Montpelliérain Alexander Grothendieck*[16] (IMAG), the mathematics laboratory of the University of Montpellier[17] and CNRS[18], located in the south of France.

Contacts

The internship will be supervised by IMAG members:

- Joseph Salmon: joseph.salmon@umontpellier.fr; http://www.josephsalmon.eu/
- Ghislain Durif: ghislain.durif@umontpellier.fr; https://gdurif.perso.math.cnrs.fr/
- François-David Collin: Francois-David.Collin@umontpellier.fr

Applicants should send a resume and a motivation letter as soon as possible.

Salary

Gross monthly salary: approx 550 Euros.
This work will be funded by the ANR CaMeLOt ANR-20-CHIA-0001-01.

Duration

The internship could last from 4 to 6 months.

Location

The internship will be located in Montpellier (Univ. Montpellier), inside the mathematics department (IMAG).

References

[1] Y. Deng, A. Kanervisto, J. Ling, and A. M. Rush. "Image-to-Markup Generation with Coarse-to-Fine Attention". en. *Proceedings of the 34th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, July 2017, pp. 980–989. URL: https://proceedings.mlr.press/v70/deng17a.html (visited on 11/16/2021).

[2] M. Downes. "Tex and latex2e". *Notices of the AMS* 42.11 (2002), pp. 1384–1391.

[3] R. Fateman. "Handwriting + speech for computer entry of mathematics". *Style, Benjamin L. Kovitz, Manning Publications Company* (2004).

[4] G. Genthial and R. Sauvestre. *Image to Latex*. Tech. rep. Stanford University, 2017.

[5] A. Jensen and H. Marklund. *Turning Equations into Latex: Pipeline vs. End-to-End*. Tech. rep. Stanford University, 2017.

[6] J. Memon, M. Sami, R. A. Khan, and M. Uddin. "Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR)". *IEEE Access* 8 (2020). Publisher: IEEE, pp. 142642–142668.

[7] R. Sharma, B. Kaushik, and N. Gondhi. "Character recognition using machine learning and deep learning-a survey". *2020 International Conference on Emerging Smart Computing and Informatics (ESCI)*. IEEE, 2020, pp. 341–345.

[8] V. Storchan and J. Beauschene. "Data Augmentation via Adversarial Networks for Optical Character Recognition/Conference Submissions". *2019 International Conference on Document Analysis and Recognition (ICDAR)*.

[16] https://imag.edu.umontpellier.fr/
[17] https://www.umontpellier.fr/
[18] https://www.cnrs.fr/

ISSN: 2379-2140. Sept. 2019, pp. 184–189. DOI: 10.1109/ICDAR.2019.00038.

[9] J. Wang, Y. Sun, and S. Wang. “Image to latex with densenet encoder and joint attention”. *Procedia computer science* 147 (2019). Publisher: Elsevier, pp. 374–380.

[10] Z. Wang and J.-C. Liu. “Translating math formula images to LaTeX sequences using deep neural networks with sequence-level training”. *International Journal on Document Analysis and Recognition (IJDAR)* 24.1 (2021). Publisher: Springer, pp. 63–75.