Image-text matching for large-scale book collections

Artemis Llabrés Arka Ujjal Dey Dimosthenis Karatzas Ernest Valveny

Computer Vision Center
Universitat Autònoma de Barcelona

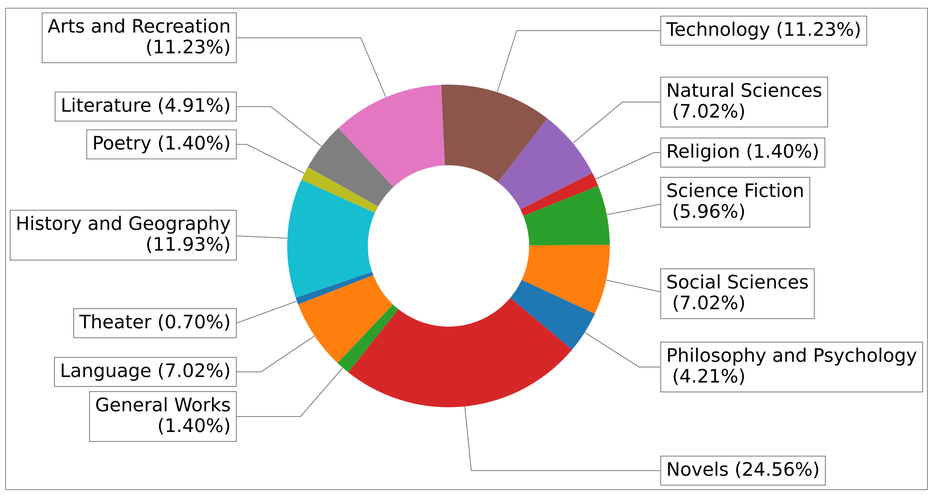
Introduction

Motivation: Automatic inventory of books from pictures of bookshelves.
Approach: Many-to-many matching between bookshelf images and a list of target texts.

Library Dataset

- 285 Images of bookshelves.
- 7536 Total number of books in the images.
- 2 target lists for many-to-many image-text matching:
  - Library List: 15k books.
  - Large list of Popular Books: 2.3 million books (including library list).

High diversity of book types and languages:



- 14 library sections.
- 7 languages: Catalan, Spanish, English, French, German, Italian, and Arabic.

Dataset Annotation:

- Annotation file with which books appear in each image.
- Amazon's Rekognition OCR: to extract text from the book spines.
- Meta's Segment Anything Model (SAM): to segment the book spines.

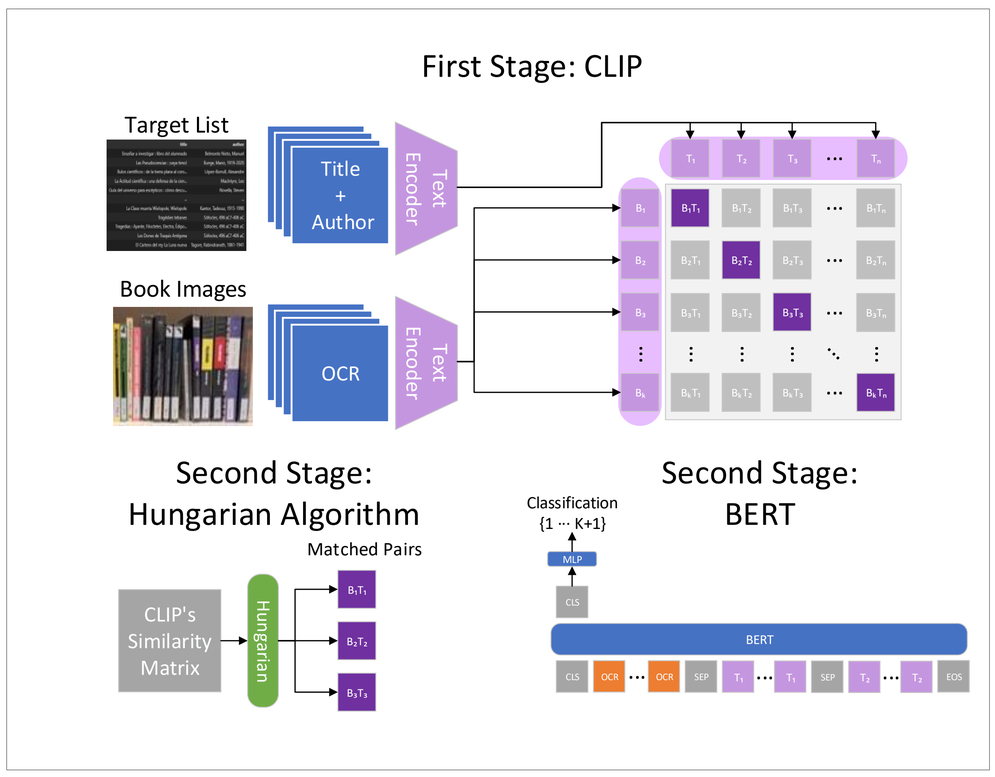
Acknowledgements: We want to thank Laura Solà from the Library of Volpelleres Miquel Batllori for the help given during the collection of the dataset.

This poster is part of the project PLEC2021-007850, funded by MCIN/AEI/10.13039/501100011033 and the European Union "NextGenerationEU"/PRTR.

Baselines and Methods

Fuzzy String Matching Baseline:
Simple baseline that uses the Levenshtein distance to compute the similarity between the text on the book spines read by the OCR with the author + title that are on the target list.

A two stage approach for many to many matching:



First Stage:

- Based on CLIP computes a similarity matrix between all books and all targets.
- Vary fast, even at large scales.

Second Stage: Two options to refine the predictions from the first stage.

- **Hungarian Algorithm:** finds the matching of one-to-one pairs of books and targets that has overall the highest similarity.
- **BERT:** For each book finds the best match from the top *K* most similar targets selected by the first stage.

Experiments

Matching Only task: Uses the already detected book spines.

Methods	library	all
String matching	0.742	**0.505**
CLIP	0.915	0.449
CLIP + hungarian	**0.918**	-
CLIP + BERT	0.824	0.310

Detection and Matching Task: Evaluates the end-to-end pipeline.

Methods	library	all
String Matching	0.573	**0.389**
CLIP	0.617	0.241
CLIP + hungarian	**0.641**	-
CLIP + BERT	0.622	0.240

Take home message

- Dataset of annotated bookshelf images that covers the whole book collection of a public library in Spain.
- Matching at scale with a target list of 2.3 million books.
- Two-stage approach for text-image matching.

Download the dataset!

mailto: allabres@cvc.uab.cat