Dealing with data issues for AI-supported Image Analysis in Cultural Heritage: concrete cases and challenges

Slides

References and links

This document: https://bit.ly/31qncRH

General (Europeana) AI links:

- EuropeanaTech AI Task Force
- Europeana Foundation Machine Learning discussion paper,
- EuropeanaTech dataset challenge
- Saint George on a Bike

BnF:

- GallicaPix
- GallicaSnoop
- api.bnf.fr
- CENL's AI cookbook for libraries
- ai4lam.org

Survey and results

Discussion & Questions - Please type your name and question here!

Question for José: you mentioned the work done on super resolution of images, which seems like one of the out-of-the-box solutions that has little to no downsides or issues, and can be used immediately on digitised cultural heritage. Are there any challenges for using this tool on a large scale? - Jolan Wuyts

Question for Antoine: could you expand a bit on the methods you used to ensure a nicely balanced selection? Was this achieved through (mostly) manual curation, or were any computational tools used? - Jolan Wuyts

Question for Jean-Phillippe: How did you organise the user testing of GallicaSnoop? How did you approach the users, and how did you collect and process the feedback or results? - Jørgen Schyberg from The National Library of Norway

Questions for Jean-Philippe: When you engage with users (human-in-the-loop), do you use Zooniverse as well? Or do you have your “own” platform for crowdsourcing? Do you have plans to develop crowdsourcing tools in your interface when outrolling GallicaPix in Gallica? Do users tend to use GallicaPix instead of Gallica? Or do they mostly use Gallica Classic? - Mette Kia Krabbe Meyer - Royal Danish Library

Question for José & Antoine: have the Europeana AI projects been mostly within Europeana’s data team, or is the EuropeanaTech community involved in some larger way? - Tom Cramer, Stanford

Question for Antoine: based on the data challenges you have cited for the St. George project, do you see updating or enhanced practices for Europeana participants to facilitate AI, not just aggregation? - Tom Cramer, Stanford

Question for Antoine Isaac: Could you give us a few examples of the weird things that happened in your dataset?
Aurélia Rostaing, Archives nationales

Question for Jean-Philippe: Could you share some practical / organizational / strategic advice about leading / mobilizing internal and external team and expertise at such a large institution to access, assemble, clean datasets and run ML experiments on datasets from in-house collections? Vincent Cellucci - TU Delft Library

Questions for Jean-Philippe: Did I understand it right that the datasets - for instance the one with soldiers (WW1) - included users and was a crowdsourcing (human-in-the-loop) project? Or was it maybe inhouse annotation? Mette Kia Krabbe Meyer - Royal Danish Library

Notes on presentations

Jean-Philippe Moreux, expert scientifique Gallica, Bibliothèque nationale de France

- GallicaPix (2017-) https://gallicapix.bnf.fr/rest?run=findIllustrations-form.xq
    - Uses an Extract, Transform and Load (ETL) pipeline (incl. IIIF)
    - Groundtruth creation (12 classes model, e.g. drawings, photos, advertisements, scores, comics, handwriting, engravings maps, ornaments, covers, blanks, texts)
    - Some cases it was simple, but others required bootstrapping
    - Evaluation: recall/accuracy, user & in-house testing, usability survey
    - Confusion matrix (90-95%). Even a 1% error is too much for a human (visually obvious)
    - Semantic indexing using IBM Watson Visualisation Recognition API (quite good on contemporary documents) => trained on pictures of **soldiers**

    - Hybrid solution good for users
- GallicaCIP (Classification of Heritage Images)
    - Uses https://www.bnf.fr/fr/mandragore
    - Data Sparsity Issue
    - iNaturalist
    - https://github.com/tzutalin/labelImg
- GallicaSNOOP
    - GallicaSnoop - Ministère de la Culturehttps://www.culture.gouv.fr › Atelier-INRIA-2019
- REMDEM - Musical scores

Antoine Isaac, R&D Manager at Europeana

Saint George on a Bike

- Barcelona Supercomputing Centre & Europeana Foundation
- Detect angels etc in art
- Challenge of selecting data sources for training
- Manual annotations, then using the first results to produce more annotations semi-automatically (Humans in the Loop)
- Challenges (of aggregating data from aggregators)
    - Availability of appropriate data
    - Provenance management
    - Link Rot
    - Rights
    - Balance of selection
    - Duplicates (aggregation)
- Can AI also help with data curation ?
    - Image similarity
    - Duplicate detection (distance between images)

José Eduardo Cejudo Grano de Oro, Machine Learning Engineer at Europeana

- Image tagging pilot; via Zooniverse
- https://pro.europeana.eu/post/introducing-our-image-classification-pilot
- Image tagging with commercial services (e.g. Google vision API, AWS Rekognition, Microsoft Azure Vision, IBM Watson Visual Recognition) - comparison with custom model
- Watermark Detection (image classification: with/without watermark)
- Super resolution: artificially increase the resolution of thumbnails in Europeana using Machine Learning
- Image similarity: useful for: recommendations, clustering, duplicate detection
    - Uses self-supervised learning
    - Image recommendation (ca. 60K images from Europeana collections)
- How can we work on this as a community?
    - Experiment sharing (e.g. via Google Colab)
    - Data Sharing (e.g. via Europeana Community in Zenodo)

- Future Plans
    - Cost-benefit analysis
    - Continue experiments and then move to deployment/scalability
    - Reporting to the European Commission
    - Development of a Python Interface
- Discussion items
    - What data issues have you encountered?
    - How you used ML to solve them?
    - Have you used commercial services?
    - Do you reuse data from other CHIs?
    - How do you share your data and experiments?