TOWARDS A NATIONAL COLLECTION

UKRI Arts and Humanities Research Council

FINAL REPORT
COVID-19 PROJECTS

Visitor Interaction and Machine Curation in the Virtual Liverpool Biennial

Leonardo Impett, Principal Investigator

Durham University | Liverpool John Moores University | Liverpool Biennial

December 2021

TABLE OF CONTENTS

Executive Summary

This report summarises the AHRC-funded Towards a National Collection COVID-19 Project, *Visitor Interaction and Machine Curation in the Virtual Liverpool Biennial*, which ran from 1 January 2021 to 31 August 2021. The project was based at the Department of Computer Science, Durham University, in collaboration with the Liverpool School of Art and Design, Liverpool John Moores University and the Liverpool Biennial.

A summary of our research project and its main questions is provided in the abstract. The aims and objective section outlines our three principal research ambitions: to prototype a different use for machine learning in virtual exhibitions (as co-curators, not search engines); to understand how visitors might interact with such a system; and to look at the bias present in the machine learning algorithms that power it. We then report details of our administrative structures (partnerships, staffing, and timetable).

Our research methodologies and results are summarised in the *Research Approach*, *Research Results*, and *Project Outputs* section. We have three principal project outputs, tied to our three research questions: an online machine curation prototype hosted by the Liverpool Biennial, ai.biennial.com, and associated open-source codebase to reproduce it on other collections; a visitor interaction dataset, which is freely viewable inside our online browser, and an associated anonymous survey which is not public; and an open-source toolkit for examining bias in the major machine learning algorithm used in the project.

The report concludes with our project's recommendations for the Towards a National Collection program, in the context of our unusual status as an AHRC project in a computer science department.

Abstract

Visitor Interaction and Machine Curation in the Virtual Liverpool Biennial was funded from 1 January to 31 August 2021. Our project started from the observation that most machine learning and artificial intelligence systems are deployed in a GLAM (Galleries, Libraries, Archives & Museum) context as either search engines, or as ways to automate cataloguing. In addition, the machine learning systems used in GLAM settings (e.g. textual or visual search engines) are almost exclusively 'uni-modal': they work with one modality of information at a time.

Instead, we proposed to use machine learning systems in a more tightly interactive setting, as a mixed-initiative system (an important paradigm for computer-human interaction in the context of artificial intelligence research). Furthermore, we used machine learning systems that translate between modalities; that turn images into texts, and vice versa. Beyond developing and launching our mixed-initiative co-curation system with the Liverpool Biennial, we have also spent time investigating the implicit bias in the most widely-used multimodal neural network, OpenAI's 2021 CLIP. Understanding bias in such networks will be an important part of using them in GLAM settings, both for mixed-initiative interactive systems like ours, and for more traditional search-engine or cataloguing-oriented systems. This led to new data (in terms of how audiences interact differently with active human-machine co-curation systems), and new research directions for digital curation, digital exhibition design, and machine learning for visual art.

Our computer-human co-curation prototype, which makes extensive use of multimodal deep learning, is online at ai.biennial.com. A special issue of the Liverpool Biennial's *Stages* journal on "Curating, Biennials, and Artificial Intelligence" was published in open access to coincide with the prototype's release. Code to reproduce the main co-curation prototype on other datasets is available on Github, alongside a separate code repository containing experiments to investigate the bias embedded in CLIP, the main multimodal deep learning network used in the project.

Aims and Objectives

Our primary research objective was to take the applications of computer vision and machine learning in the GLAM sector away from the “search engine” model, and instead explore alternative models, suggesting curation as a fruitful and productive metaphor for machine learning in museums and galleries. We were also interested in how this took us away from the virtual exhibition model of online visitor interaction; unlike in a traditional online exhibition, no two visitors to our Virtual Liverpool Biennial saw the same set of works in the same order (as they co-curated it with a machine learning powered system).

A secondary objective was to use this new prototype to collect data about how visitors interact with a system of this kind. This data can then be compared with data from a previous ‘virtual’ Liverpool Biennial, which had used a very different kind of online interaction - the Biennial’s *Minecraft Infinity Project* in 2016 (curated by the co-investigator of this project).

Finally, a further secondary objective was to investigate the biases and implicit assumptions of the computer vision and machine learning models we had been using; and whether an interactive co-curating framework might give us a space for uncovering these biases to online visitors, in a way which has traditionally been difficult to do for computer vision / machine learning powered search systems or automatic metadata generation.

Partnership Structure

The project was based at the Department of Computer Science, Durham University, where the Principal Investigator Leonardo Impett is assistant professor of Digital Humanities. The Co-Investigator Joasia Krysa is full professor of Exhibition Research at the Liverpool School of Art and Design, Liverpool John Moores University. We also had support from Durham University's Advanced Research Computing division.

Our major project partner, unfunded by the project but offering resources both in cash and in kind, was the Liverpool Biennial. This included cash support for paying the outsourced web development of the frontend of the project, which was undertaken by the MetaObjects studio (Andrew Crowe and Ashley Lee Wong), with design by Sui Lam. The Liverpool Biennial also offered in-kind curatorial support, and web hosting at ai.biennial.com, a subdomain of the Liverpool Biennial website.

Staffing Structure

Researchers

Postdoctoral Research Associate, Department of Computer Science, Durham University

Dr Eva Cetinic

Responsibilities: development of main machine learning systems; data cleaning from Liverpool Biennial; design direction

---

Research Software Engineer, Advanced Research Computing, Durham University

Mark Turner

Responsibilities: rewriting and cleaning code into reusable and re-readable state; collating and developing experiments into open source package

---

Research Associate, Department of Computer Science, Durham University

Hiu Yuen

Responsibilities: data scraping and writing scaleable experiment scripts for CLIP testing

Investigators

Principal Investigator

Dr Leonardo Impett, assistant professor, Department of Computer Science, Durham University

Responsibilities: technical supervision, development of the overall system design, writeup and presentation for digital humanities and technical research sectors

---

Co-Investigator

Prof Joasia Krysa, professor, Liverpool School of Art and Design, Liverpool John Moores University

Responsibilities: coordination with Liverpool Biennial; dissemination and interfacing with GLAM sector; writeup of results for GLAM sector

Revised overall programme

Timeline (project months)	*Description*	*Modified with respect to plan?*
January 2021	Project start	Delayed (planned start in August; funding decision made in December)
March 2021 (M3)	Recruitment of PDRA	Delayed partly due to offer letter (mid-January)
June 2021 (M6)	Working prototype of machine curator experiment	-
August 2021 (M9)	Final release of machine curator experiment	-
September 2021 (M10)	Funded project end PDRA end	-
January 2022 (M12)	Final publication of research findings	-
January 2022 (M12)	Hybrid (physical-virtual) exhibition	Cancelled due to delayed start
June 2022 (M18)	Informal survey of impact	-

Research Approach

Our research was conducted by a strongly interdisciplinary team: including two scholars with a background in the humanities (Kyrsa, Turner) and three with a background in engineering or computer science (Cetinic, Impett, Yuen). We were keen not to pigeonhole individual research questions or Taylorise our research process - and therefore involved all researchers in key research decisions and findings on both technical and humanistic questions.

We worked very closely with the organisers of the Liverpool Biennial, including meeting weekly with Joasia Krysa (our project co-investigator and head of research of the Liverpool Biennial) and regularly with Sam Lackey (Liverpool Biennial director). Although the frontend web design and implementation was done by a commissioned agency and funded by the Liverpool Biennial, we worked together with the designers (Andrew Crowe, Ashley Lee Wong, Sui Lam) to brainstorm and develop various interaction paradigms for the co-curation system in the first four months of the project. The basic idea behind our final interaction system, which heavily involved cross-modal learning, was conceived by Eva Cetinic, and refined iteratively by the whole research group. This involved interfacing several machine learning systems (image captioning, generative adversarial networks, image-text networks, etc) with the image and text archives of the Liverpool Biennial, which are freely accessible online[1]. The pipeline for applying these systems was designed and implemented by Eva Cetinic; the code was then cleaned by Mark Turner, in order to create an open source repository which other scholars or GLAM institutions can re-use, modify or expand[2].

Because our machine curation prototype incorporated data collection (both in terms of logging individual curatorial preferences, and an optional survey at the end), our data collection phase started concurrently with the release of the Biennial prototype. Once the backend work for this was finalised (circa month 5), we were able to proceed with an analysis of the biases inherent in the most widely-used of the models in our system, OpenAI's CLIP (Contrastive Language-Image Pretraining)[3]. We used various methods to probe CLIP, specifically investigating the way in which it related to works of art, and whether (or how) it has embodied a 'canon' of visual art from its wide training dataset (which is made up of very large images with captions downloaded from across the internet). These included: zero-shot classification (can it more easily recall material, style, or subject?) and image generation (can it more easily generate famous images than lesser-known ones; and if so, which images can it recall?). These questions have a wide impact for other uses of computer vision across the GLAM sector (including as search engines and in automatic metadata generation), since CLIP greatly surpasses previous models in its ability to capture visual information in artworks, and is already being used for the automatic classification of digitised works of art[4].

[1] https://biennial.com/archive

[2] https://github.com/DurhamARC/machine-curation

[3] https://openai.com/blog/clip/ ; Radford, Alec, et al. "Learning transferable visual models from natural language supervision." arXiv preprint arXiv:2103.00020 (2021).

[4] Conde, Marcos V., and Kerem Turgutlu. "CLIP-Art: Contrastive Pre-Training for Fine-Grained Art Classification." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

Research Results

Computer Vision and Non-Figurative Visual Art

To date, the majority of research papers in computer vision applied to visual art address questions inherent to figurative images. For instance, gesture detection[5] and gaze estimation[6] rely on depicted human bodies; object detection[7] relies on depicted objects; and iconographic classification[8,9] rely on classification schemes like *Iconclass*, built around early modern Christian figurative art.

We based our machine learning system on data from the Liverpool Biennial Archive; in which most images were not of this kind. Not only were the works not directly figurative, but they were also performance-based, installation-based, or temporal. There is a much more partial relationship, in other words, between the images and the artworks.

The Biennial archive also presented new types of information, not normally present in collections of early modern paintings. Most importantly, there was a significant amount of text written by the artists (or in partnership between the artists and curatorial staff): the title of the artworks, and texts (circa 500 words) explaining the biography of the artist, their artistic practice, the characteristics and relevance of this particular work, etc. The decision to use multi-modal networks, which can deal with both visual and textual information, was thus made in response to the different qualities of contemporary art data.

Our final interactive system is an assemblage of five different neural networks:

1. An image captioning network, which creates a one-sentence description of input images
2. A keyword extraction network, which extracts ~15 keywords from a ~500 word text
3. An image embedding network, which calculates the visual similarity of two images
4. An image-text embedding network (CLIP), which computes text-image similarity
5. A generative adversarial network, which generates photorealistic images; used alongside network 4 (CLIP), it can generate realistic images given a textual input prompt.

The navigation system combines information from across these networks. Navigating through the experiment, visitors are presented with a triptych of images and texts and with three possible directions to explore. Placed in the centre is the source artwork, AI-generated image on the left and a heatmap overlaid on the source image on the right (Figure 1).

In the centre, deep text networks are used to extract the most salient keywords from the source artwork's descriptions (which can be found on the Liverpool Biennial website). Navigating in this direction, we use visual and textual links – the visual links (the similarity of the artwork source photographs) and the textual links (the similarity of the keywords from the artwork descriptions) are combined. This is a version of the *visual similarity metrics* used in search and recommendation engines we see on the internet today.

---

[5] Marsocci, Valerio, and Lorenzo Lastilla. "POSE-ID-on—A Novel Framework for Artwork Pose Clustering." ISPRS International Journal of Geo-Information 10.4 (2021): 257.

[6] Madhu, Prathmesh, et al. "Understanding compositional structures in art historical images using pose and gaze priors." European Conference on Computer Vision. Springer, Cham, 2020.

[7] Gonthier, Nicolas, et al. "Weakly supervised object detection in artworks." Proceedings of the European Conference on Computer Vision (ECCV) Workshops. 2018.

[8] Milani, Federico, and Piero Fraternali. "A Dataset and a Convolutional Model for Iconography Classification in Paintings." Journal on Computing and Cultural Heritage (JOCCH) 14.4 (2021): 1-18.

[9] Cetinic, Eva. "Towards Generating and Evaluating Iconographic Image Captions of Artworks." Journal of Imaging 7.8 (2021): 123.

On the left (in pink) is the AI-generated image, first encountered on the landing page of the project. Navigating left, visitors reach the artwork with the most similar generated image in terms of colour, form, or texture. These generated images are created only from the titles of the original artworks – nothing else. The two works are therefore connected through the visual similarity of their (textual) titles.

On the right is an AI-generated description of the photograph of the source artwork. These are the deep network's best guess at what is going on in the image, using the image alone, without any textual information. For instance, Sonia Gomes' fabric sculpture *Timbre*, leads the AI to generate the description: "a person wearing colourful clothing is sitting on a stool". Above the AI-generated description, you will see a heatmap overlaid on the original image: this is an indication of the points of the image that the AI considers important for generating that description. We believe that this is amongst the first applications of Explainable AI technology in the GLAM sector. Navigating in this direction leads you to the artwork with the most similar description, using textual similarity alone – the two works are connected through the textual similarity of their (visual) appearance.

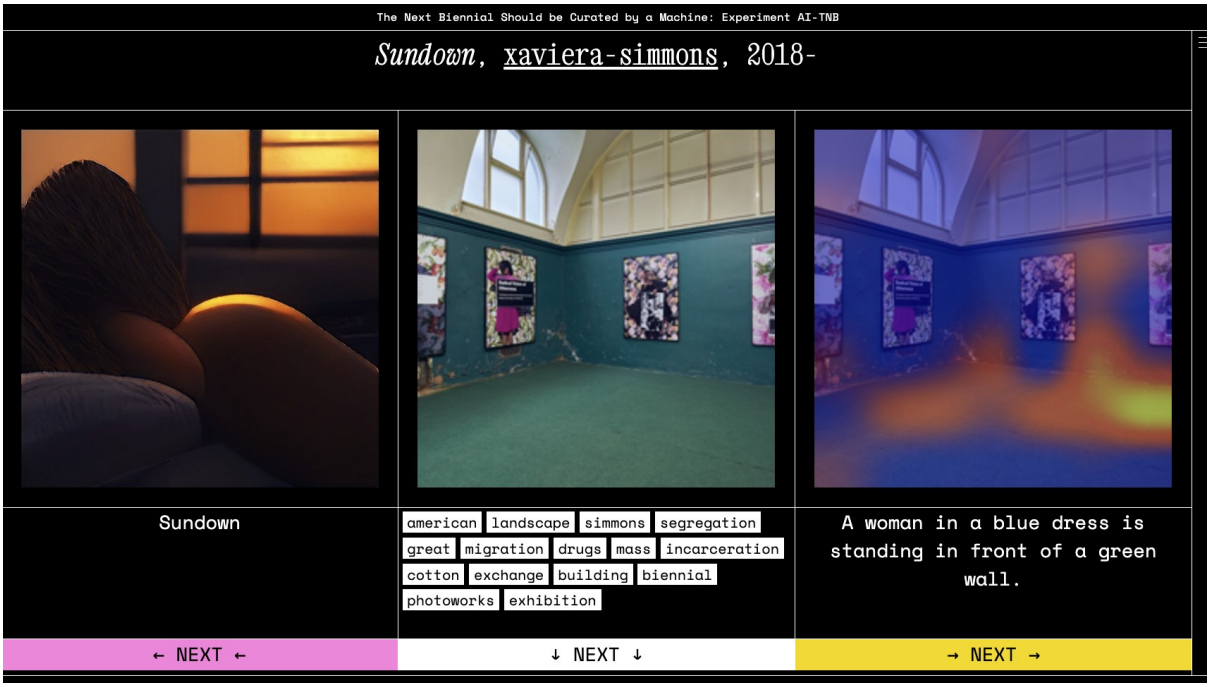

Figure 1 - screenshot of the main navigation pane of ai.biennial.com, showing Sundown by Xaviera Simmons through different computational lenses

As visitors navigate the project, they create their own paths through the material, each such journey becoming a co-curated human-machine iteration of the Biennial saved to the project's public repository. This data is available to browse through the online system (*Figure 2*); and along with the optional survey, it forms the basis for the time-sensitive data collection component of this project.

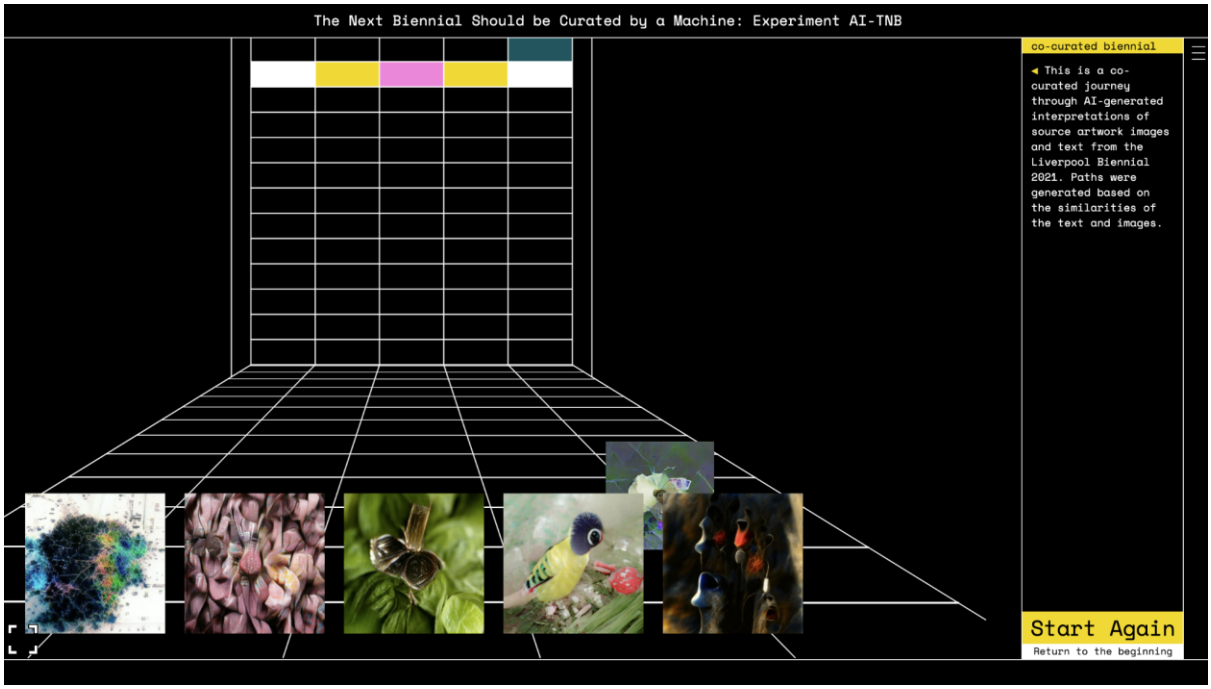

Figure 2 - exploration of previous co-curated biennials

As highlighted previously, a major research concern was the way of seeing[10] of the machine vision algorithms used in the project; especially the text-image algorithm CLIP, used to generate the machine-generated images (such as the image on the left of Figure 1). The fact that CLIP can be reversed (with the help of a generative adversarial network) into an image-generating, as well as image-classifying, tool allows us to study its implicit visual bias, memory and logic more precisely. GLAM applications of CLIP (and similar networks) must ask: does CLIP have a 'canon' of visual art? If so, what is it, and how is it memorised or processed?

Using publicly available visual art datasets (WikiArt and Web Gallery of Art), we attempted to identify these potential visual biases through three experiments:

1. Zero-shot classification. Is CLIP better at remembering some kinds of categories (e.g. the type of scene - still life vs. landscape - versus the material or technique)? Note that this is directly relevant to the task of automatic metadata generation.
2. Iconographic zero-shot retrieval. Can CLIP encode iconographic information; and if so, which kind of iconographic information does it compute most accurately (is there a bias towards Western or Christian art, for instance)? This corresponds directly to textual search or automatic tagging.
3. Image generation. Can CLIP (used with a generative adversarial network) generate images of extant artworks; and if so, does it generate more convincing images of well-known works? What might 'well-known' correspond to in this context?

[10] Azar, Mitra, Geoff Cox, and Leonardo Impett. "Ways of machine seeing." *Springer AI & Society* (2021): https://doi.org/10.1007/s00146-020-01124-6

Figure 3 - Pieter Brueghel the Elder's The Harvesters*, Oil on Wood, 1565, Metropolitan Museum of Art, public domain (left); VQGAN and CLIP image produced given the prompt: "The Harvesters by Pieter Bruegel the Elder" (right)*

The codebase behind these experiments is currently on Github, and the results are currently in the process of being written up. However, preliminary results indicate that CLIP *can* reproduce well-known works of western art remarkably well (see Figure 3); essential characteristics about their style and iconography (though perhaps not their composition) have effectively been memorised by the network. This is not the case with lesser-known artists or images. The nature of the relationship between an image's canonicity and the degree to which it is encoded in the CLIP network is still being investigated; however, CLIP's visual canon - being trained on a large quantity of automatically-downloaded internet data - is not an unmediated translation of the canon of western art history. For instance, we would expect CLIP to reproduce a well-known bronze statue when given the prompt "David di Donatello". The result is shown in Figure 4, below: a statue in marble instead of bronze is more likely due to a confusion with Michelangelo's David than Donatello's earlier (and far less well-known) marble statue of David. The green reptilian head, meanwhile, seems to be a product of the Teenage Mutant Ninja Turtle also called Donatello.

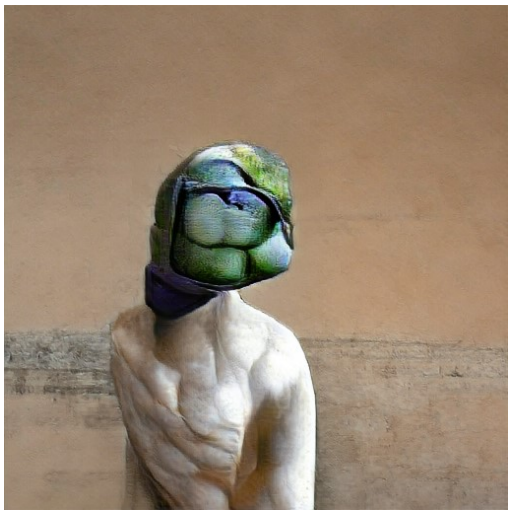

Figure 4 - "David di Donatello", as reproduced by CLIP and VQGAN

Project Outputs

Machine Curation Online System

Our machine curation system was launched in partnership with the Liverpool Biennial. Although the data on which it relies has been generated by sophisticated machine learning systems, the processing is offline, which means that the website can be served as static files; giving it a degree of longevity.

https://ai.biennial.com/

Machine Curation Dataset

Data collected from our machine curated biennials is currently collated live on a server: we collect both interaction paths (how people have moved through the exhibition; which exhibition they have co-curated) and optional survey responses. The survey is anonymous and does not contain personal data. This will be published in open access when the Liverpool Biennial domain is terminated (i.e. when we stop gathering new data). At time of writing we have over 125 individual interaction paths.

Machine Curation Toolkit

We have prepared an open source toolkit so that other collections, museums or researchers can reproduce our machine curation experiments on their own data, and so that anyone can read and understand the workflow behind our prototype. The main script is packaged into a Jupyter Notebook, *machine_curation.ipynb*, which can be run by those without access to specialised hardware on cloud computing services for free (e.g. Google Colab).

https://github.com/DurhamARC/machine-curation

https://colab.research.google.com/github/DurhamARC/machine-curation/blob/master/machine_curation.ipynb

Bias in CLIP Experiments

We have also prepared a separate open source code repository for our experiments regarding bias in the main neural network we have used, OpenAI's CLIP.

https://github.com/yuenhy/stapler/

Publications

Our project was the focus of an open access special issue, "Curating, Biennials, and Artificial Intelligence", of the Liverpool Biennial Stages journal. This includes an editorial and two articles on our project (Krysa & Impett, "The Next Biennial Should be Curated by a Machine"; Impett, "Irresolvable Contradictions in Algorithmic Thought"): https://www.biennial.com/journal/issue-9

Krysa and Impett have a further article in press on the project: "The Next Biennial Should be Curated by a Machine", *AI and Humanities* (ed. Freddy Paul Grunert), European Commission Joint Research Centre, 2022Cetinic, Yuen and Impett have a paper in draft on the results of our bias in CLIP experiments, to be submitted early 2022.

Recommendations for the programme

Our research project had the fairly unusual status of contributing to an AHRC-based project from within a computer science department. The vast majority of the costed research time on the project (our PDRA, RA and our PI) were within this department, and a portion of time was also budgeted to a Research Software Engineer from Durham University's Advanced Research Computing.

This allowed us to go further on the technical side than an applied science project might have been able to. Our research brought cutting-edge machine learning and computer vision techniques - including multimodal network guided image generation, and interpretability heatmaps for image captioning - to the GLAM sector for what we believe to be the first time. Both of these techniques help us (both as users and as scientists) to better understand the biases and assumptions of our machine learning systems, which we believe is essential both for public trust and for intellectually critical use of machine learning systems in a GLAM context.

Our first recommendation to the program, therefore, is to bring in *computer scientific* knowledge - rather than simply technical capacity - into the design and delivery of humanities infrastructure programmes like *Towards a National Collection*. This is a different - and complementary - form of expertise to the mixture of best practices, software engineering and library and information science expertise commonly provided by digital humanities groups. This is especially true for fields like machine learning, deep learning and computer vision, which are highly specialised and rapidly changing subfields of computer science.

Good models for cross-disciplinary collaboration of this kind are furnished by bioinformatics and medical imaging; both of which make significant use of new developments in machine learning and computer vision. The case of medical imaging is particularly similar to computer vision in the humanities; with small datasets, expensive expertise needed for labelling, images which are frequently non-photographic (or fundamentally different in nature to "generic" image datasets like ImageNet), and research questions motivated fundamentally by disciplinary expertise (i.e. from medicine or the humanities). Like in medical imaging, then, computer vision for the humanities would benefit substantially from building its own set of methodologies and technologies, rather than merely applying the state-of-the-art from conventional (photographic) computer vision. In the case of bioinformatics and medical imaging, theoretical and applied research is funded across councils (e.g. by EPSRC, BBSRC, MRC and NERC), and interdisciplinary training (e.g. PhD programmes) is relatively mature.

At present, this is not the case for the digital humanities in the UK (with the possible exception of the UKRI Digital Economy theme); and there are significant research areas that would fall into the (unfundable) gaps between AHRC and EPSRC; for instance, developing *new* neural network architectures for humanities and GLAM applications that avoid the eurocentric and anachronistic biases of CLIP. Instead, most relevant tracks in the digital humanities (NEH-AHRC, UK-Ireland collaborations in DH, *Towards a National Collection*) are exclusively AHRC-led, meaning technical research such as ours is limited to the reimplementation or slight modification of existing algorithms. Given the uniqueness of their data and research questions, computer vision and machine learning applications for the GLAM sector, for *Towards a National Collection*, and for the humanities more generally, might instead benefit considerably from purpose-built models for the humanities; as well as technically-led investigations into the fairness, accountability and transparency of machine learning and computer vision tools in the humanities.

Contacts

Leonardo Impett, Assistant Professor, University of Cambridge

li222@cam.ac.uk

Joasia Krysa, Full Professor, Liverpool John Moores University

J.M.Krysa@ljmu.ac.uk