FUTURS FANTASTIQUES

Vendredi 10 décembre 2021

Grand Auditorium

Daniel van Strien, Andrew Longworth & Catherine Cronin, Flyswot : garden-variety machine learning applications

Catherine Cronin

Hi everyone, gonna jump straight in.

What do all these images have in common?

We can see that they are images of collection items, but it's not obvious what else they have in common. If I pause on this image here, this is a screen grab from a digital object platform called Digitised manuscripts (http://www.bl.uk/manuscripts), we can see that this image that's displayed has got an image label “f-i-“r which -- what does that mean to most people? Maybe not much, but to people in the know, this label stands for fly sheet one recto and we can see that this is an image which is not a fly sheet it's of a Torah mantle. So the thing that all these images have in common is on this digitised manuscript website, they have all been given an image label describing them as a flysheet but they are not flysheet so we call these images fake flysheets. Digitised manuscripts was a platform designed for a particular project, it has a limited set of image labels that fitted the requirements of that project. When the platform went live, curators at the British library were very excited about this platform and they wanted to publish their manuscripts and material to this platform even though their physical items and images did not meet the label requirements of the platform. So on this platform we have lots of images that are described as a flysheet but they're not actually images or flysheets. The image label you see on the platform is tied to the image file name. We have a new viewer at the library called the universal viewer: this has got a much more flexible and large vocabulary of image labels, that means when we migrate content from the old platform Digitized

manuscripts to the new platform we have the opportunity to give images that are incorrectly labelled as flysheets a more appropriate label. And here we have some screen grabs of our universal viewer and the images you're seeing now typically would have been called flysheets in our old platform Digitized manuscripts. So here we have edges of a book and we can give them appropriate labels, we have a loose leaf that has second foliation in roman numerals we can give it a more appropriate label, and we have an unfolded case image again we can give it a more appropriate label, and we have a three-quarter view of that we can give a more appropriate label. So not only are we going to be migrating content from our old platform to our new viewer, we're also taking 20 years worth of digitized content ingesting it for preservation into our digital library store and making all of that content available in the universal viewer. We have a mountain of content and in that mountain of content we know we have many images that have been labelled as a flysheet image but they are not an image of a flysheet. So how can we identify them easily to give us the opportunity to give them the correct label on our new platform which is the universal viewer?

[Daniel van Strien]

So this is where computer vision comes in as a potential approach to solving this problem and specifically we wanted to create a workflow for creating a computer vision model that could be updated with new training data. I guess the other key point of this model and this workflow is that we wanted it all to be fairly boring because the goal of this particular piece of work was to fit into an existing workflow and set up the computer vision and machine learning pipelines in a very pragmatic way. So as part of that we used tools like DVC data version control to create a reproducible pipeline that when given new data could be retrained and produce a new model that then we could apply to the flysheets we want to detect. Next slide please. So once we have this trained computer vision model which is able to detect whether an image it's shown is a fake flysheet or not, we deploy it as part of a command line application and that was chosen as a kind of approach to deploying this particular computer vision model because it fits well within existing workflows and sits within existing tools that people who are going to be using this tool are familiar with. And so what this command line tool does is it's given a directory of images to look at and it recursively looks through all of those directories and subdirectories and looks for images which

have a flysheet in the file name and then it produces a prediction whether that flysheet is actually a flysheet or not. And that prediction is put inside a CSV report which is later used by the person that's running that tool. (And next slide). This is just a quick screenshot of the flyswot application and so although we are using the command line interface we're still trying to give the user of the tool quite a lot of information about the model and this is something we want to continue to build on so that the kind of computer vision underpinning the flyswot tool isn't hidden from the user. And I'll pass over to Andrew.

Andrew Longworth

Thanks Daniel. So I'm just going to speak a little bit more about how we fit this software into our general workflow, so at first when we built this tool we thought it might be a magical tool that would instantly label any image that we had and would save us a lot of manual time and effort visually inspecting images, but we very quickly realized that, because we unfortunately – in inverted commas – have a very large and varied collection with lots of subtle distinctions across collection areas, that this automatic labeling wasn't going to be possible and a bit more than that: one collection areas labels might actually differ from another collection areas labels what are very similar items so manual intervention or verification was always going to be needed in some form or another. But what we primarily wanted as Daniel mentioned was a tool that would tell us the difference between a real flysheet and something else so a tool that would just point us in the right direction, telling us where we needed to make a manual intervention. Then in terms of from a practical point of view of integrating it into the workflow we figured it would make sense to integrate this tool with some other software we already used, so we built some Power BI reports to look directly at the CSV results which just make it much easier to understand and consume. So one of the instant headline results we got was the total number of items that we'll need looking at more closely in a batch of images. So how many items in this batch of hundreds items have obvious fake flysheets or images where the model is not really sure it's a flysheet so we'll need double checking. This is very useful because it helps us to understand how much work will be needed to change the labels in this batch of images and then we can make decisions around resource allocation and prioritization of work. (Next slide please. Catherine please). So here's an example of some of the visuals from the Power BI report and another good thing

about this integrating it with tools we're to use Power BI, is it allows us to immediately focus in on the specific items with fake flysheets, so rather than looking through a long list of file paths in a CSV and trying to pick them out, power BI just instantly shows us a list of just the items and their specific tips which need re-labeling or investigation. Next slide. And then an added bonus about the Power BI integration is that it quickly highlighted some unexpected results and not immediately obvious predictions. For instance here you can see cases where the first prediction, perhaps we didn't say but there are two predictions made in the model, first prediction is that it says it's a flysheet and then the second prediction says that it's not a flysheet and so obviously this is something we look to understand a little bit further and we did look to understand a bit further and that helped us to refine the model and its threshold. So using these tools we already have sort of fed back into the process and helped us to improve the accuracy. Next slide please. So finally I guess in conclusion we just wanted to highlight how this machine learning software is just one of the many tools that we use as part of our adaptable workflow, and we do use it on a regular basis. Manual intervention is still required as I mentioned but it's actually desirable not just because of the subjective nature of the items in our collection like I mentioned but also because I think this is quite important this semi-automatic process that we employ is great for building trust and confidence in the library in machine learning. Because that's not necessarily something that most curatorial colleagues are familiar with or comfortable with machine learning but we're hoping they will become more so in the future so this is a great way of sort of bringing them into that the workflow user and getting them to use machine learning. Next slide please. Catherine. That's all from us really. Thank you very much.