ML in the Newsroom

OK, so now let's try to understand machine learning by the different types of problems that it can solve. I'm going to arrange this by data type, so machine learning for images, versus text, versus audio, and so on.

So let's start with talking about machine learning for photos. So in the last section, we talked about how we might want to do this with classification, classifying photos of cats and dogs and labeling them as dogs or cats. This is a super useful tool, actually, vision classification models. And for example, it's really popular in the medical industry, where maybe we have X-Ray scans of lungs, and we want to identify pneumonia or not. For all different types of diseases and medical scans, it turns out that, often times, machine learning models are more accurate than trained practitioners. So very popular in medicine. And of course, also useful in data journalism.

So let's talk about some common machine learning mission tasks. One thing you might want to do is identify what's in an image, in general. So here's a picture of Luna Park, which is an amusement park, and we've used a vision API to identify what's in this picture. So we get these labels: landmark, 96 percent. That means the model is 96 percent confident that this is a picture of a landmark. It also thinks it's a place of worship, which is not quite right, but it sort of understands that this is some grand, temple-like structure. Maybe it's a tourist attraction.

By the way, these labels come from this tool called the Cool Cloud Vision API. I'll talk about it much later in the Tool Section, but this is just a common sort of type of task you might want to do with a vision model.

You might also want to identify faces. And if there are faces, see if there's any emotion within those faces. So this model doesn't find any emotion. This face, it says it finds headwear. But you can imagine maybe using this. Suppose somebody is giving a rally. And you get a nice big swooping shot of an audience, and you want to know what percentage of the audience is happy, versus angry, versus sad, or whatever.

Of course, no machine learning model is perfect, so you have to make sure you double check and make sure you're OK with errors. But emotion detection is a pretty common task.

And OCR. This is just a text extraction. So maybe you have signs, as in this photo. Or maybe you have scans of PDF's, or forms, or handwritten notes. You can use machine learning to extract those images of text into actual text.

So how do people actually use these tools for images in the newsroom? Well, one interesting project was a recent collaboration that Google did with The New York Times. So The New York Times has existed long before the digitalization of photos. So they had this enormous archive of physical pictures, as you can see, here. And on the back of these pictures would be all these sorts of notes from editors, and the photographer, and the reporter, explaining where the photos should go and what it is about. But it was very hard for New York Times to deal with all these photos because they were all physical. How are you going to sort through them?

So with Google, they scanned their photos and then used our vision tools to identify, for example, what was in the pictures, in order to extract the text, and then to ultimately make this sort of searchable digital archive. So machine learning is really very good for organizing this sort of unstructured photo text data into something that's more manageable.

The New York Times also worked on this sort of other fun project. They used facial recognition to build a bot where you could text a picture of a member of Congress, and it would tell you who it thought the person was and with what confidence. So the idea is that there are so many members of the House of Representatives. How are you gonna remember them all? Well, if you see someone you think you might recognize walking down the street, you take a picture, and The New York Times tells you if it thinks its a member of Congress. This is a facial recognition tool.

And finally, another interesting project that used machine-learning provision was a piece by a Texty. So they wanted to identify illegal amber mining, which was destroying environments. People illegally mining in order to sell. So they collected all of these satellite pictures of different regions, where they knew contained and didn't contain illegal amber mines. And they used experts to identify which ones were the illegal mines and so forth. And they used all these pictures to train a machine-learning model to look at new satellite images and to try to automatically classify whether the site contained illegal mining. And as a result, they were able to make this interactive map, where they could point out where they thought all legal mining was occurring.

So that's an overview of machine learning with images. What about machine learning for audio data? You probably have a lot of audio data in the form of interviews. And wouldn't it be nice if a computer could come along and transcribe your interviews for you? That's certainly a great intersection of machine learning and journalism. I think it's something that we've worked with newsrooms to try to make happen. To try to help with the entry transcription.

But you can also use speech-to-text transcription for lots of investigative reporting. So, for example, if you have hours and hours of videos or voice footage and you want to analyze it, well, it's not really useful in those formats. Wouldn't it be so much faster if you could have that data in text format like transcriptions?

So, for example, there's a lot of work done by Kalev Leetaru. He's a reporter that's published a lot in Forbes. And he's used transcription and different vision AI tools to analyze television, weeks and weeks of television, and wrote a bunch of pieces for Forbes on it. I've included them in the extended reading section.

So, for example, this isn't audio. This is vision. But he looked at film from all these different news stations ABC, CBS, NBC, and so on. And he identified all the faces using our vision tool and tried to calculate how many were expressing joy. He found that ABC was the happiest, and PBS was the saddest. He also used this to identify what a new station was showing a Trump tweet. It was able to tell that CNN showed the most Trump tweets, and PBS unsurprisingly showed the fewest Trump tweets.

So, again, a lot of different things that you can do when instead of having a video or audio file, you instead have text that you can analyze. Speaking of text that you can analyze, let's talk about machine learning for text data.

Before we talked about classifying photos of cats or dogs, but of course we could also classify blocks of text. So, for example, sentiment classification. That's whether a text is saying something positive or negative. Maybe we want to know what's the best airline. So we go online, and we scrape all these tweets mentioning different airlines. Jet Green and Ulta. And we label a bunch of them ourselves, and then we can build a machine learning model that automatically identifies positive and negative tweets.

Or maybe we have a bunch of articles that we want to automatically categorize, so we want to sort the electronics articles from the politics articles, from the cooking articles, say. You can also, of course, do investigative reporting with text classification.

So one of the most impactful and interesting pieces I've seen is from the L.A. Times back in 2014. They wanted to see if they could identify crimes that had been misclassified. So, for example, when a crime occurs, the LAPD classifies it as being a violent crime or not a violent crime, and I guess The L.A. Times wanted to double check. So they collected a bunch of descriptions of crimes, and they had human beings label them as being violent or nonviolent. And they trained a model then that could take a description of a crime and predict: was it a violent or nonviolent crime? Then they looked at thousands of. LAPD's descriptions of crimes, and found that actually a lot of violent crimes involving stabbings, whatever, were being mislabeled as nonviolent. And so in a way, artificially deflating the violent crime rate in L.A. So this awesome peace used machine learning to sort of uncover that.

Now, finally, let's talk about machine learning for tabular data. This is just the sort of data that you would find in a spreadsheet or a database. Numbers, categories, things like this.

So BuzzFeed wrote a piece analyzing tabular data a couple of years ago. It was a really interesting piece, where they wanted to be able to identify hidden spy planes. So they used data from a site called Flight Radar 24, which has all sorts of information on all flights that occur like a flight's altitude, and how long it lasts, and the positions, and stuff like this. All this information about flights. And BuzzFeed knew that certain flights were spy plane flights and some weren't. And they used this to build a model that, given all of this data about a flight, could predict whether it was a spy plane or not.

What they found was really fascinating. They found, for example, that there were some planes that were supposed to be tracking terrorism in Africa, but there were actually flying over U.S. cities. And they found, for example, planes that were tracking drug cartels on the border. And lots of things that work were sort of unexplained to them when they wrote the piece. They were able to to learn more about them.

So the model would say this is a spy plane, and then the reporters would try to verify. However, the model made mistakes a lot. For example, it consistently recognized skydiving planes as spy planes because they also have these weird, loopy trajectories. So there's a great lesson to be learned here, which is that all machine learning models make errors, and sometimes they make them consistently, like consistently misclassifying skydiving planes.

So it's important when you're a reporter to understand how to work with these models, and when they make predictions that say something damning like this is a secret spy plane, that you as the reporter have to go and then verify that plane using traditional reporting.