Module 2: Sorting images with fast.ai

John Keefe: The first step in any machine learning project is loading in our data. So, let's jump over to our notebook and get started with that.

Research Doc Google dot com and click on GitHub and Quartz.

And this time we're going to be using D.D. sorting images with fast.ai, and we've gone through the plan here pretty well. We just need to get our setup going, which includes making sure that we set our runtime type to GPU to take full advantage of that powerful Google computer and its GPU.

Next, we want to load the fast day I program into our notebook. So, we're going to go ahead and do that. Skip this next cell. That doesn't pertain to us, but this one does. Loading in the library is loading in the first day I library and we're ready.

Next, we need to get the data. And the data really is all of those images - the circling maps and the non-circling maps. So, we're going to pull those down. And I'll just show you, by listing in the directory, here is a directory called Data Slash Choppers. And there are two folders in there - one is full of circling maps and one is full of not circling maps. So that's all set.

We're going to set that as our data path. So, we can use that in many locations. And now we're going to talk about this block here. This is what fast.ai calls a data block. And it's really important to get all of these images and all of these labels and everything together in the right form and fast.ai makes it really easy to do it with this data block.

So, this line we're going to skip right now. It's about some transformations we can do on the images to randomize them a little bit. But we're not going to do that right now. We'll do that in a couple of chapters. But here the data that we're going to feed into this data variable, first, we're going to say go get those images from a folder in our data path. So, we set our data path up here. It's going to go in and check in that chopper's directory for all of our images.

And then, it's going to take those images and it's going to split them randomly. And this is actually super important. It's going to split them into a training and validation set.

And let's step aside for a minute and just show you the importance and difference between a training and validation set.

Remember - we're using images to train a model, right? And they're labeled images. They're labeled as circling or not circling. And so, the computer is trying to learn which patterns of pixels represent circling and which patterns of pixels represent not circling. And it basically guesses, guesses and guesses a bunch of times and checks its work.

So what we do is we split all of these images that are in the data bunch into two sets. One is called the training set and one is called the validation set. And you can think of the training set as like homework or maybe like a practice test. Right? So, here are a bunch of images. *“Computer guess which ones are circling, which ones are not*”. And you get it gets scored on that and it adjusts and says, OK. OOPS, I made a mistake. I made a mistake. And you can imagine it going and going and getting really, really good at the practice test, at the homework.

But then the moment of truth happens. Right. It has to be presented with images it hasn't seen before to see how well it would work in the wild. That's the validation set. That's like the exam. Right? So, you don't want to be overly good at just answering the practice test questions. You need to also be good at determining whether these this new set of pictures is circling or not circling. So that's training and validation in our fast.ai Data block.

We're not actually separating them just like this, just like the last 10 or 20 percent we're actually going through and randomly picking out throughout the data bunch which images we're going to use for the validation test, which ones will be held out and used as the exam during the repeated epics of training.

OK. So, back here in the notebook, we've got that split, taking care of. The model needs to know how to label these images. And we're going to say just use the name of the folder for the label on the image. OK? That transforms. We're going to skip that for now. But this is important. All of these images are of size 600. And because they're maps, I'm going to keep them at size 600. This last line is about packaging it all up into what's called the data bunch. And this is the batch size is going to be 16. That's just how many images we're going to throw at the processor at a time. We're just going to say 16. That's going to be just fine for us. So, I'm going to play this cell.

And that's done. And now let's take a look at what we have.

OK, so this is just looking at our data. We haven't done any processing or anything except to put the correct labels to go with each one. You can see these are the folders I put them in. This map was in the not circling folder, so it gets a not circling label. This one down here, you can see it was circling. So, I put this map in the circling folder and now it's got the labels circling. So now the computer knows or at least has together the maps and their labels. And here we're going to make sure that we only have two classes just to be super confident about that.

That's it.

We're all set now that we have the data loaded. Next, we'll actually train our model and see if we can use it to sort these images.