SDS PODCAST EPISODE 620 FIVE MINUTE FRIDAY: OPENAI WHISPER: GENERAL-PURPOSE SPEECH RECOGNITION

(00:05):

This is Five-Minute Friday on OpenAI Whisper.

(00:27):

With their latest model called Whisper, the revered A.I. research firm OpenAI has claimed that speech recognition with machines is now effectively a solved problem. Their justification for that claim is that their new Whisper model approaches the accuracy and robustness of humans. This adds yet another model to the catalog of breakthrough models that OpenAI has released recently such as GPT-3, you can hear all about GPT3 in Episode #559, as well as DALL-E 2, which you can hear about in Episode #570, and OpenAI Codex, which you can hear about in Episode #584.

(01:07):

Ok, so back to Whisper. One of the challenges holding machines back from approaching human-level speech recognition like Whisper has has been acquiring sufficiently large amounts of high-quality, labeled training data. “Labeled” in this case means audio of speech that has a corresponding text associated with it. With enough of these labeled data, a machine learning model can learn to take in speech audio as an input and then output the correct corresponding text.

(01:36):

To go one level deeper, a common modern way to do this is to have a so-called “encoder” portion of the machine learning model to convert the audio into an abstract, high-dimensional, numeric representation. Then subsequently, the so-called “decoder” portion of the machine learning model converts that abstract, high-dimensional, numeric representation into a sequence of natural language text.

(02:02):

More specifically, Whisper leverages an encoder-decoder Transformer, similar to the famous OpenAI GPT-3 that inputs and outputs text. Whisper has different applications from GPT-3 because, while Whisper does output text, it takes in audio as an input.

(02:21):

Another similarity to GPT-3 is that Whisper excels at so-called "zero-shot learning" wherein you request a model to perform a task it may have never encountered before but it nevertheless conjures up a sensible output. As examples, OpenAI have demonstrated Whisper to excel at diverse tasks such as: language identification, appending phrase-level timestamps, multilingual speech transcription, and translation to English from other languages. About two thirds of the data Whisper was trained with was in English, so English translation is the only target translation language currently supported.

(02:56):

So, we've covered the Whisper model architecture and learned about its capabilities, but it's critical to point out that what enabled this model architecture to be so effective and have all these capabilities is the large, high-quality dataset that OpenAI collected for Whisper to be trained on. Historically, one of the challenges with speech recognition has been acquiring large amounts of high-quality, labeled training data. An older, well-known speech recognition model called wav2vec for example, used a self-supervised approach, where a machine learning model trains itself to learn one part of an input from another part of the input, thereby meaning we only need to provide the model with unlabeled audio of speech. This method could easily be scaled up because we can easily find lots of unlabeled speech audio, but it results in only being able to train the encoder portion of the encoder-decoder structure needed for outputting text

that corresponds to the audio input. A trick used in previous speech recognition models was to train the encoder on a huge unlabeled dataset and then train the decoder by fine-tuning the model with a smaller, labeled dataset. This decoder fine-tuning trick resulted in many models that performed well on recognizing speech similar to that contained in the smaller, labeled dataset but the model wasn't robust to recognizing other manners of speaking.

(04:20):

In order to train a more robust, generally applicable model, OpenAI collected 680,000 hours of labeled data from the Internet — making it the largest dataset ever created for training a speech recognition model — and they trained their encoder-decoder Transformer architecture end to end with this huge dataset. This massive dataset contained data labeled for a wide variety of tasks — such as multilingual speech recognition, speech translation, spoken language identification, and voice activity detection — which is how Whisper came to be so effective at zero-shot learning across a wide range of tasks.

(04:58):

Furthermore, the diversity of audio contained within the massive dataset — from different environments, recording setups, speakers, and languages — is what makes Whisper so robust. Whisper may not perform as well as other models that specialize in the specific speech recognition benchmark called "LibriSpeech", for example, but because Whisper is not fine-tuned to any specific dataset, it is markedly more robust: It makes half as many errors as those specialized models do across many diverse speech datasets.

(05:29):

Does Whisper now sound interesting and you'd like to check it out? I did and I was seriously impressed; it flawlessly transcribed everything I said to

it even if I tried to mumble or speak with unusual inflections. Thanks to Hugging Face, you can try Whisper's transcription capabilities out yourself too, right now. We've included a link in the show notes.

(05:48):

Thanks to Shaan Khosla, a data scientist on my team at my machine learning company Nebula for inspiring this Five-Minute Friday episode on Whisper today by providing a summary of the Whisper paper via his Let's Talk Text Substack newsletter. He uses the newsletter to provide a weekly easy-to-read summary of a recent key natural language processing paper and so you can subscribe if that's something you're interested in — we've provided a link to Shaan's Substack in the show notes as well.

(06:16):

Ok, that's it for this episode. Until next time, keep on rockin' it out there, folks, and I'm looking forward to enjoying another round of the SuperDataScience podcast with you very soon.