Zoe Winkworth

Final Project Write-Up

For my final project, I decided to create an Alexa Skill to voice enable an app with visual components by including vocalized image descriptions from a computer vision API. This project was partially motivated by the current state of image description on the web for visually impaired and blind users. The work that I read described that many images on websites are not adequately described, and that many images are not described at all, leading to a disappointing browsing experience for many users (Petrie et al., 2005; Nganji et al., 2013). Studies also show that creative audio descriptions significantly increase user engagement and enjoyment of media (Walczak et al., 2017). The most important descriptions users wanted for images were of objects, buildings, people in the image so I decided to focus on providing that information (Petrie et al., 2005).

I had to do a decent amount of research to find an available API for an app with significant visual components that used the version of authorization protocols supported by Alexa Skills Kit, and eventually landed on the Reddit API because it fit these requirements. I used the Alexa Skills Kit, with a custom interaction model defined in the Alexa Developer Console and endpoints written in Nodejs 8.10 defined in a Lambda function hosted in Amazon AWS. For the image description component, I used the Amazon Rekognition API also hosted in AWS. I also used AWS Cloudwatch, AWS CloudFormation and Postman for debugging.

I took an iterative approach to creating the skill, beginning with a very simple skill that simply fetched the top post from a hardcoded test-based subreddit with default sorting and passed the title to Alexa to read. I then added additional text-based subreddits and customizable sorting before adding image processing. After adding support for two image-based subreddits, I added additional browsing capability by retrieving multiple posts and keeping track of posts users had seen before to provide new content when users asked for posts from a given subreddit multiple times. I then attempted to add video processing and account linking.

I encountered a number of challenges throughout the development of this project. As briefly touched on earlier, many of the APIs that I was initially hoping to use were either not publicly available or did not used an updated OAuth version. Because I needed to access the API from the Alexa Skills Kit, I was limited by the compatibility requirements of that system. Luckily, I was able to choose a computer vision API that was hosted by Amazon, which was much more compatible with Alexa Skills Kit. The only change I needed to make to accommodate the AWS Rekognition API was to move away from the Beta Alexa Hosted endpoints option, and create my own Lambda function on AWS in order to have more flexibility in how I handled external dependencies.

I also ran into some limitations of the Amazon Rekognition API capabilities. There are some very strict limitations of the format of images that are supported by the API (only JPEG or PNG) and the image bytes are required to be provided in a base64-encoded string. Since I don't have a lot of experience with image processing or file formats, it took me some time to figure out how to accomplish this. Additionally, I discovered that the API would not accept image URLs beginning with "https", only "http". So without being able to determine how to change that programmatically, I ended up doing some string manipulation to remove the offending "s". Similarly, I discovered that the video processing functions only supported MP4 file formats, so I was unable to add features to support "/r/gifs" as originally planned.

I also ran into a lot of problems trying to implement the account linking necessary to allow users to "vote" or comment on posts via Alexa. Navigating the complex series of steps necessary to adhere to the OAuth guidelines was a non-trivial problem. OAuth implementation seems to vary from platform to platform, and while Reddit API documented their implementation relatively clearly, the Alexa Skills Kit documentation was somewhat scattered and often outdated. The main issue that I faced was in troubleshooting. Since both sides of the authentication process were owned by third parties, it was difficult to gain insight into any issues other than the generic "Unable to link your skill at this time". Even after setting up an HTTP proxy service in AWS CloudFormation, I found that I was not familiar enough with authorization protocols to identify any issues in the logs and Alexa support did not respond.

I tested several scenarios on the Alexa simulator in the Alexa Developer Console. I made sure to test each of the subreddits that I added support for, all of the custom sorting options, and I tested that asking for the same subreddit multiple times results in a new post each time. At the moment, I do not support the ability to increase pagination, so after a number of requests the posts will loop. Also, if you switch subreddits and come back, you will see posts starting from the top again. In some cases, the pictures in a post from "r/pics" may be in an unsupported format, in which case Alexa will return an error message. My code also supports cancel, exit, repeat and help.

While I tried to be mindful of the type of information I was including in my image descriptions (basic keywords, facial features of any faces in the picture and any text) I found that my descriptions were often not relevant and were sometimes long enough to be tedious to read or hear. I tried to add more conditionals to reduce information that the API was not confident about, or that I could tell would always be irrelevant. For example, I require that most information have a confidence score of over 90% before the system includes it in the description. The only exception being that if the system is unsure of the gender of a face, it prefaces the initial comment about the face with "I'm not certain, but". I also added the constraint that information about facial hair is only included in a description if the system is confident that the face is male.

Despite this, I found that problems still came from the Rekognition API having no sense of focus or relevancy. Especially with the feature to detect text in an image, the API would often include text that was on items in the background, or clothes worn by people in the picture. It was also bad at recognizing words, so the text found was often gibberish or random letters. Additionally, I found that simply describing salient objects or people in a picture were insufficient for actually understanding the context of the picture or the actions happening in it. I think a lot more work can be done in finding more targeted and effective ways to computationally describe the content in images to improve accessibility of apps and devices.

References

Petrie, Helen, Chandra Harrison, and Sundeep Dev. "Describing images on the web: a survey of current practice and prospects for the future." *Proceedings of Human Computer Interaction International (HCII)* 71 (2005).

Nganji, Julius T., Mike Brayshaw, and Brian Tompsett. "Describing and assessing image descriptions for visually impaired web users with IDAT." *Proceedings of the Third International Conference on Intelligent Human Computer Interaction (IHCI 2011), Prague, Czech Republic, August, 2011*. Springer, Berlin, Heidelberg, 2013.

Walczak, Agnieszka, and Louise Fryer. "Creative description: The impact of audio description style on presence in visually impaired audiences." *British Journal of Visual Impairment* 35.1 (2017): 6-17.