Azure Price Estimate on Whale Identification Systems Project

App Service

This component is for hosting the web application of the systems project. The web application is a Plotly Dash application (developed in python) to allow users to upload images of whale flukes for identification. The web application will not do any inferencing, it will solely take images uploaded by the user, resize them, and then pass them to an Azure Function (explained below) for inferencing. Then results returned by the model in the Azure Function will be displayed in the web application. Resizing images to 256x512 resolution is not only a requirement for the model, but it ultimately reduces the memory footprint of each image to about 0.4 MB. Additionally, the max number of images a user will be able to upload to the web application is 25 images. So, the maximum total amount of RAM needed for the images is 10 MB. With this image cap, there is no need for a high amount of RAM for the app service. The lowest tier of this service provides 1.75 GB of RAM and this should suffice for the use-case of the systems project.

Storage

The storage component is used for storing uploaded images for retraining purposes. Once a user uploads an image, it will be stored in Block Blob Storage. The selected capacity for the Block Blob Storage is 1 TB, which is more than enough for storing the uploaded images since each stored image is ~0.4 MB in size. Furthermore, the size of the entire dataset, with a total of 15k images, used for training was about 5 GB in size. So, the total amount of images we can store with this component is 1 million 256x512 images.

Azure Functions

Azure Functions are a serverless compute service provided by Azure meaning that the code ran by this service is not sitting on a server that is up 24/7, it is code that is only ran when triggered. Additionally, you are only charged for time spent for the code to fully execute. These charging guidelines are commonly referred to as “Pay as you go”. The model used for inferencing (i.e predicting identifiers of whales) will be “hosted” in an Azure Function.

The Azure Function will need to be able to hold the Pytorch model in memory and the images that will be fed through the model when triggered. Moreover, the memory footprint of the model is 44 MB and as stated before, the memory footprint for 25 images is 10MB. The minimum amount of memory provided by this service is 128 MB, which is more than enough. The Azure Function will not have access to a GPU for inferencing for cost-saving purposes. Ultimately impacting the inference time to be longer than ideal. The compute time was set to 60,000 milliseconds because it is expected that inference can take up to 1 minute when the max number of images allowed are fed through the model. Lastly, the amount of executions per month was set to the default, 10000 executions/month, which is highly overestimating. Fortunately, the service still satisfies the free tier with these parameters so using this service for our use-case will incur no costs.