FIRE: THE FIRST-YEAR INNOVATION & RESEARCH EXPERIENCE

CLOUD COMPUTING

Automated Cloud Identification Comparing VGG16 And Google Teachable Machine

Anthony Louie, Matthew Stauffer, Kenny Cheung, Ethan Pak

Introduction

The foundation of all machine learning models boils down to the data used to train it. In order to optimize the performance of said models, the data used for training should be properly cleaned[1]. For NASA's GLOBE Cloud, a civilian-submitted image database, certain images contain things such as obstructions or were non-cloud images. In order to have proper data that will be used for training of models such as classification models[2], we need to separate images that do not meet the standard for ideal training into ready images for training and flagged images that should not be used using 50% obstruction as our baseline.

Tools & Data

- VGG-16: A type of Neural Network architecture that is used to create a supervised machine learning model.
- Google Teachable Machine: a web-based, supervised machine learning model that uses a provided dataset of images
- NASA GLOBE Cloud: An observer-based cloud dataset of cloud images with associated information such as time and location. Images are collected by citizen scientists.

Figure 1. Left: an example of a flagged image where there was >50% obstruction. Right: bn example of an image with minimal obstruction that we would've used for training

Literature

1. Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016). Data Cleaning. Proceedings of the 2016 International Conference on Management of Data. https://doi.org/10.1145/2882903.2912574
2. Huertas-Tato, J., Rodríguez-Benítez, F. J., Arbizu-Barrena, C., Aler-Mur, R., Galvan-Leon, I., & Pozo-Vázquez, D. (2017). Automatic cloud-type classification based on the combined use of a sky camera and a ceilometer. JGR: Atmospheres, 122(20). https://doi.org/10.1002/2017jd027131

Results

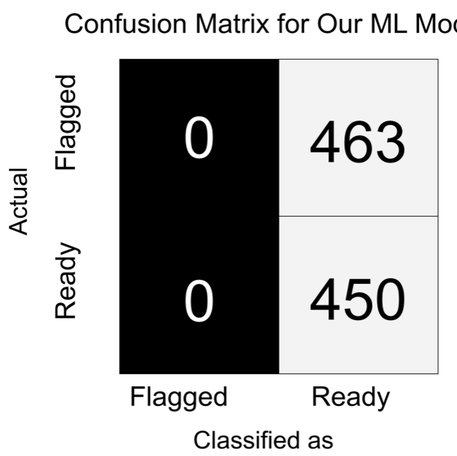

Figure 2. The confusion matrix for our VGG16 model. The results show that the model identified all images as ready, meaning we have a flaw.

In order to have some kind of baseline to assess our VGG16 model's performance, we made a teachable machine to compare to. The same sets of training and testing data was used. Metrics calculated by dividing the number of images classified by the total number of images actually in that category.

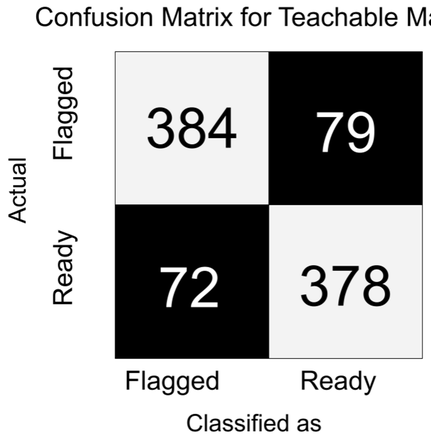

Figure 3. This is the confusion matrix for the Teachable Machine. The results show us that this model was able to correctly identify a majority of the ready images, but was not as strong at identifying the flagged images.

Teachable Machine Performance

- Flagged Accuracy: 83%
- Ready Accuracy: 84%
- Overall Accuracy: 83.5%

Discussion

After looking at Figure 2 and Figure 3, we can see that the overall accuracy of the Teachable Machine model (83.5%) is not same as the model using VGG16 (~50%). Since we used the same data set to train and test both models, it's reasonable to assume that both models would perform somewhat equally, that's not the case.

It should be noted that the VGG16 model had only 5 epochs of training while the Teachable Machine model had 50 epochs of training. An epoch is a full training process of a version of the model being trained. Each epoch is made up of a number of steps. Generally, more epochs and more steps per epoch usually results in more accurate measurements.

Our research question and overarching goal was to determine whether we could create an automated system to flag cloud images that we determined to be unfit to be used for future incorporation in machine learning systems.

One of the things that is important to note about our VGG16 results is that, despite the model recognizing all of our testing data as Ready, the prediction values of the model, which cannot be shown by the confusion matrix, do change. The two confidence values hover around 45% (confidence that the image is Flagged) and 55% (confidence that the image is Ready), with a variance of around ±2%.

Overall, we would hold off on deploying our VGG16 model for data cleaning as the output would not filter the dataset. If you want a very basic form of data cleaning, you could use the Teachable Machine, but we don't recommend using any of them.

Anthony's Current Work

Currently, I am approaching the same problem but stepping away from the CNN approach described early. I am instead using image thresholding to calculate cloud fraction from NASA GLOBE Cloud observations to see if the upward image from the observation can represent the overall observation. The upward image is free of obstructions, so it is ideal for training/testing cloud classification models. And if it is representative of the entire observation than key information stays.

Figure 4. This is an example of an upward image from an observation from the NASA GLOBE Cloud Database. There are no obstructions and the image has a clear depiction of cloud and sky.

Thresholding Techniques

Figure 5. This is the original, unfiltered, image from the NASA Globe Cloud Database. This is an example of an image with high complexity such as different cloud intensities and numerous obstructions on many parts of the image.

Figure 6. This is the image from Figure 5, but masked using adaptive thresholding with a 55 pixel area to find a mean threshold and a constant of 5 to show major obstructions.

Figure 7. This is the image from Figure 5, masked using otsu thresholding.

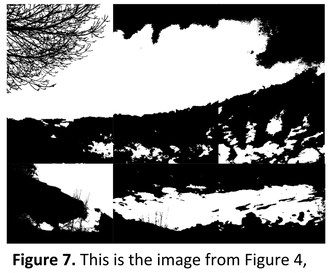

Figure 7. This is the image from Figure 4, but segmented into 9 portions first and then masked using otsu thresholding.