Machine Learning Engineer Nanodegree

Capstone Proposal

Craig Houston
October 19, 2018

Domain Background

Air pollution has become the world's single biggest environmental health risks. Nine out of ten people around the world breathe air containing high levels of pollution, and air pollution kills 7 million people each year[1]. One of the most dangerous types of air pollution is fine particulate matter (PM2.5) which is small enough to enter deep into the lungs and pass into the bloodstream. It has been linked to a host of diseases including lung cancer, stroke cardiovascular disease and adverse birth outcomes.

Awareness of the dangers of PM2.5 pollution is slowly increasing, and over the last decade many countries have taken steps to monitor concentrations of this pollutant. However, installing reliable measurement equipment can be expensive and many poor countries and rural areas lack the resources or technical expertise to purchase or maintain reliable measurement equipment.

I have personal experience living through extreme air pollution in Beijing for two years, and raising a young child in SE Asia, where seasonal burning causes high levels of pollution every year. My desire to protect my son's health lead me to build a network of low cost sensors in Thailand to raise awareness of the PM2.5 pollution problem. Through our network of internet connected sensors, I have access to high resolution pollution data from the city of Chiang Mai, Northern Thailand. I also have a database of almost 4000 photos taken hourly of the city over a period of 162 days spanning the 2018 smoky season. My goal for this project is to try and train a convolutional neural network (CNN) to correctly categorize air pollution using images alone. If successful this work could have applications in poor countries and rural areas where installing a sophisticated air pollution sensor is not feasible, but using a low cost camera or smartphone is feasible.

Problem Statement

Monitoring of air pollution in poor countries and rural communities is severely lacking. This problem is particularly acute in Africa, where only 8 out of 47 countries are recording particulate matter pollution.[2] Ideally, pollution measurements should always be carried out using certified monitoring equipment, but the cost of this equipment and

---

[1] http://www.ccacoalition.org/en/news/world-health-organization-releases-new-global-air-pollution-data

[2] https://www.who.int/phe/health_topics/outdoorair/databases/en/

lack of skills to properly maintain it are significant barriers. For such communities where no monitoring is available, an alternative approach using low cost cameras to capture images and computer vision to identify pollution could be a good first step to inform communities about their local air pollution. This knowledge could support the targeted installation of specialist pollution sensors in areas identified as having high pollution levels.

Proving this concept (and solving the above problem) requires a suitable set of images which contain a range of pollution conditions, each with a corresponding pollution classification. Based on the author's own measurements in Chiang Mai, Thailand, both of these critical input datasets are available. Therefore the problem is quantifiable, measurable and replicable.

Datasets and Inputs

The following datasets and inputs will be used:

Pollution images: A time-series of 4,000+ photos taken hourly from an elevated position over the city of Change Mai (Figure 1) . Images are 1640 x 922 pixels in dimension, captured using a 1080p Raspberry Pi Camera[3]. The time span of images cover the period from 16 February 2018 to 22 July 2018. This coincides with the annual burning season which typically occurs during February - April.

Figure 1: Example of images take over Chiang Mai during polluted (left) and un-polluted (right) conditions.

Pollution data: Two types of PM2.5 (fine particulate matter) sensors co-located with the camera recorded continuous data throughout the period of image capture. These are both light scattering laser sensors, one was an AirVisual Pro[4], the other was a Purple Air PA-II[5] . Further detail of the measurement characteristics of each sensor are presented below.

---

[3] https://uk.pi-supply.com/products/raspberry-pi-camera-board-v1-3-5mp-1080p

[4] https://www.airvisual.com/air-quality-monitor

[5] https://www.purpleair.com/sensors

Sensor	Sampling frequency	Relevant data channels
Air Visual Pro	Every 15 minutes	PM2.5 concentration, temperature, humidity
Purple Air	Every 80 seconds	PM2.5 concentration, temperature, humidity

By matching the timestamps from the photos and the pollution data, each photo will be assigned a PM2.5 AQI category and index value based on guidelines US Environmental Protection Agency's Air Quality Index (AQI)[6]:

AQI Category	Index Value	PM2.5 concentration (µg/m3)
Good	0 - 50	0.0 - 12.0
Moderate	51 - 100	12.1 - 35.4
Unhealthy for Sensitive Groups	101 - 150	35.5 55.4
Unhealthy	151 - 200	55.5 - 150.4
Very Unhealthy	201 - 300	150.5 - 250.4
Hazardous	301 +	250.5 +

The pollution dataset exhibits a bimodal distribution with a relatively long tail, as shown in Figure 2.

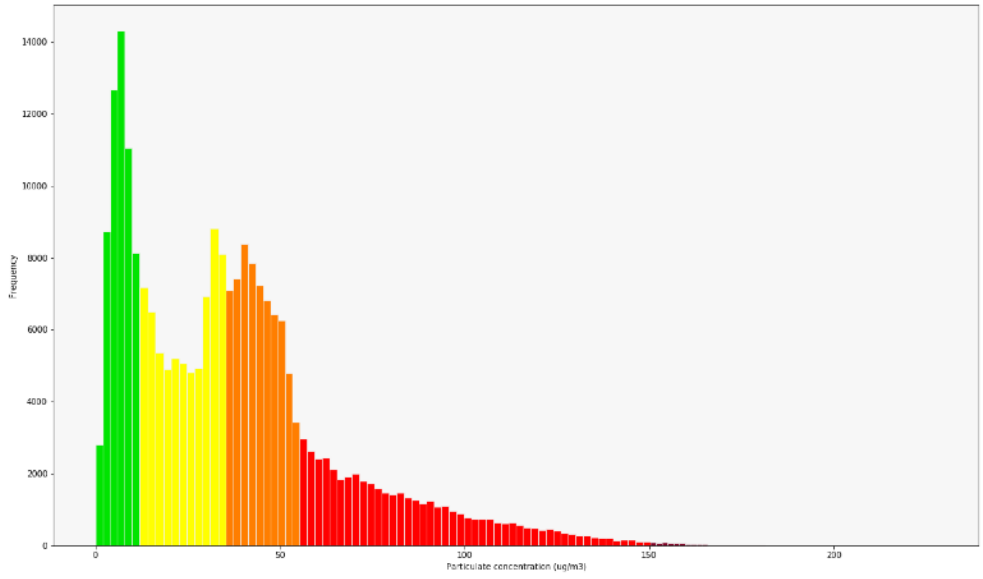

Figure 2: Histogram of pm2.5 data from PurpleAir sensor

[6] https://www.epa.gov/sites/production/files/2016-04/documents/2012_aqi_factsheet.pdf

This first mode of the distribution corresponds to episodes of relatively clean air, while the second mode corresponds to burning season pollution. The distribution of the target classes is an important consideration when choosing the most appropriate evaluation metric. The table below shows that the first 4 classes have a relatively similar weight in the data set, while the very unhealthy air is quite rare at this location.

AQI Category	Proportion of dataset (%)
Good	24.1
Moderate	28.3
Unhealthy for Sensitive Groups	27.5
Unhealthy	19.7
Very Unhealthy	0.4
Hazardous	0

The goal of the CNN will be to correctly classify the AQI category of each image. A further step that might be explored is whether the CNN can also accurately predict the index value, but achieving this level of precision is considered unlikely.

Transfer learning dataset: Transfer learning using the Places365[7] dataset is proposed in order to improve the performance of the CNN and reduce training time. In contrast to the ImageNet database, Places365 images are more tailored to training CNNs for scene recognition and is considered a better match for this application.

Solution Statement

The proposed solution to the the problem of identifying air pollution in the absence of suitable ground measurements is to train a CNN to correctly classify pollution from images. Transfer learning will be applied using a model pre-trained on the Place365 data set. This is considered the best approach since the available dataset of pollution images is relatively small, and there is significant overlap with the image types in Places365. A classification layer with six nodes matching the AQI categories will be trained, while freezing the weights in the pre-trained model layers. The performance of the solution can then be measured by testing how accurately the CNN classifies the pollution categories in a withheld group of test images.

Benchmark Model

The most basic benchmark model, and lowest performance level that the model should surpass is the random assignment of AQI categories to each image. There are 6

[7] http://places2.csail.mit.edu/download.html

categories in the AQI classification system, but data captured at the test site shows no measurements within the hazardous category, so we will limit ourselves to 5 categories. Random selection would therefore be expected to return an accuracy of approximately 1/5 or 20%.

Additional investigation into the use of CNNs for pollution categorization did not reveal much prior work in this field, with the documented research papers all behind paywalls. This makes it difficult to identify a standard benchmark that can be objectively compared to the solution, if one even exists. Therefore, a single layer CNN without transfer learning is considered an appropriate benchmark.

The author proposes to compare the model with his own classification of images into AQI categories. Intrinsic human subjectivity means this is not an ideal benchmark model, but it can at least serve as a more stringent approach than random guessing. Furthermore, medical studies using a doctor's interpretation of images provide a precedent for this type of benchmark. If the model can perform better than human interpretation of the images, then it can offer value to those trying to understand air pollution in the absence of ground based measurements. The author's benchmark may not be entirely representative of the general population since he has experience comparing images and pollution data, but it would be a worthwhile test.

Evaluation Metrics

The distribution of the air pollution data shows some instances of very unhealthy air, but the frequency of these events is quite rare. The sparse data in this category will present a challenge for the CNN to extract and learn the features present in images with very unhealthy air. It also means selecting a simple accuracy evaluation model would be problematic since this would likely result in a model that never predicts the very unhealthy category. Clearly this is not ideal, since we want the model to alert us to high pollution conditions.

Two solutions to the above are proposed. The first is to us an F1 score evaluation metric in order to balance the need for precision and recall, while penalizing the model for failing to identify the very unhealthy category. This may still fail due to the limited data available in this category. The second approach will be to merge the unhealthy and very unhealthy categories into a single classification category. F1 score will be the preferred metric in this second case, however a simple accuracy metric may also be sufficient since this new classification split will result in fairly even distribution of data points in each category.

Project Design

Below is the proposed workflow, including a discussion of implementation considerations and possible strategies to explore:

Preparing the data

Firstly the data inputs must be cleaned and processed. This ensures that the input signal to the CNN is formatted correctly and free of aberrations that may adversely impact its ability to achieve the stated goal. This process will include the following:

Image data

- Removing images taken during nighttime hours, as these provide no valuable image data.
- Remove timestamp text at the top of the images.
- The camera used to capture the pollution images faces west, therefore several images capture the sun directly near sundown. These images will be removed as the high exposure drowns out any signs of pollution.
- Resizing images to reduce total image size and format to a square aspect ratio.
- Since images have been taken on a static camera, each photo is positioned identically. This should help isolate the changes in the image conditions (pollution level). It also means that image augmentation (scaling, offset, rotation, mirroring) should not be required.

Pollution data

- Check for data timestamp consistency - this check is very important when trying to match the time varying pollution data with the corresponding image.
- Isolate measurements that coincide with the exact timestamp of the image. This way we get the pollution level right at the instant the image was taken, and not a data value which has been averaged over a longer time period.
- Remove any erroneous data points or outliers. These are common with measurement devices in the field and should be carefully removed to avoid skewing the dataset. Having data form two separate air pollution sensors will help to identify such erroneous data points.

Building the model

There are several CNN architectures that may work well for this pollution image classification problem. As a starting point, several of the more recent CNN architectures shown to perform well on this type of problem with be tested. In particular MIT has made available a number of CNNs pertained on the Places365 dataset, these include VGG-16, ResNet50, ResNet152 and GoogLeNet.

The loss function to be used will be categorical cross-entropy, as is appropriate for multi-class categorization problems.

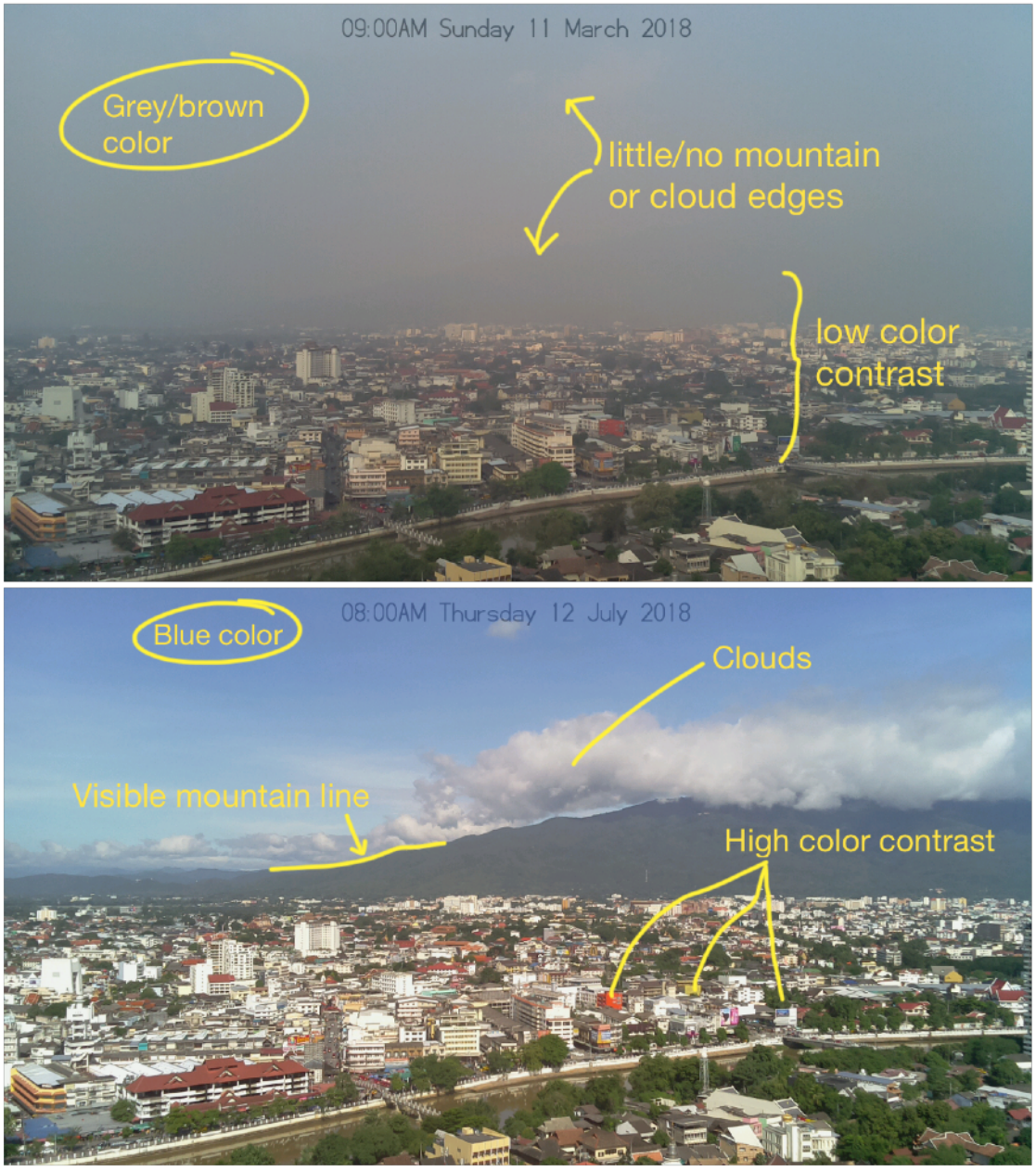

Figure 2: Visual features characterizing polluted (top) and non-polluted air (bottom).

For the model to be successful it will need to correctly identify the key differences between images with and without pollution. The annotated comparison below shows the types of features that a human eye would use to make this distinction between a polluted day (top) and a clear day (bottom):

These visual differences may be more subtle that the differences between say, a mountain and a lake, that a CNN pre-trained on the Places365 data would typically classify. However, since the Places365 dataset is comprised of many of these type of landscape images, it may still be able to pick up the differences.

It is expected that the model will have the most difficulty distinguishing between pollution and high humidity conditions, as these are visually similar. In this case the use of humidity data recorded by the sensors may be beneficial.

Another model set-up question regards the effectiveness of model layer depth for this application. The model does not have to classify hundreds of different objects, rather it has to learn how the image of the same scene changes with the pollution level. This may suggest that a very deep model is not necessary.

Since color is one of the important characteristics that indicates the presence of air pollution, a 3-dimensional RGB filter will be used. The most appropriate convolution window size is unknown at this point, so this will be tested in a range of 2x2 up to 5x5.

Finally different activation functions will be tested to see which can best activate the subtle differences between images of polluted and non-polluted air.

Evaluation

The model discussion above highlights several parameters that will be investigated in order to optimize the performance of the CNN. Each will be evaluated on how accurately the resulting model can categorize pollution levels from a withheld group of test, as well as resources needed for training.