

Predicting Air Pollution Levels from Satellite Images Using Deep Convolutional Neural Networks

Arnav Bansal

Introduction

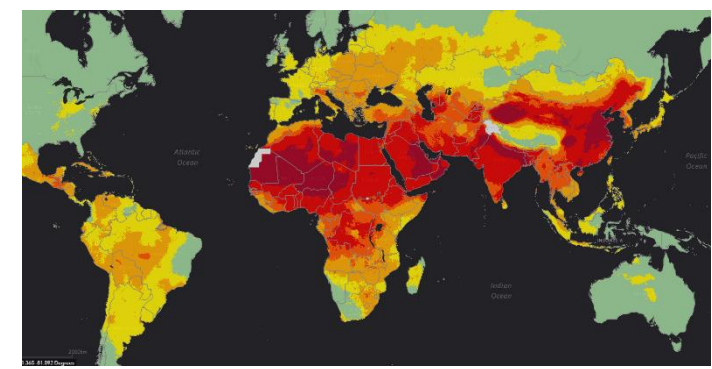
Air Pollution

Leading environmental threat to public health

Linked to 6.5 millions deaths worldwide in 2015

92% of people worldwide are exposed to unsafe air pollution levels

Exposure causes heart disease, lung disease, stroke, and premature death



Air Pollution Forecasting

1. Provides health recommendations that keep people safe, including young children, senior citizens, and asthma patients
2. Enables governments and corporations to take measures to reduce global air pollution

Existing Methods

Traditional Air Pollution Sensors

- Advantage: Forecast air pollution with fair accuracy
- Disadvantages: Expensive, susceptible to damage and failure, limited in geographic coverage, lack a presence in underdeveloped countries

Spectroscopic Techniques & Filter-based Gravimetric Methods

- Advantage: Forecast air pollution with fair accuracy
- Disadvantages: Complex, not scalable

Social Media-based Methods

- Advantage: Forecast air pollution with fair accuracy
- Disadvantages: Not scalable to underdeveloped countries that have primitive social media, errors in sensor data accumulate to these models

Existing Image-based Methods

- Advantage: Forecast air pollution with fair accuracy
- Disadvantages: Limited in geographic coverage as images are ground-level, not scalable to underdeveloped countries as many images are obtained from social media, errors in sensor data accumulate to these models

Other Methods:

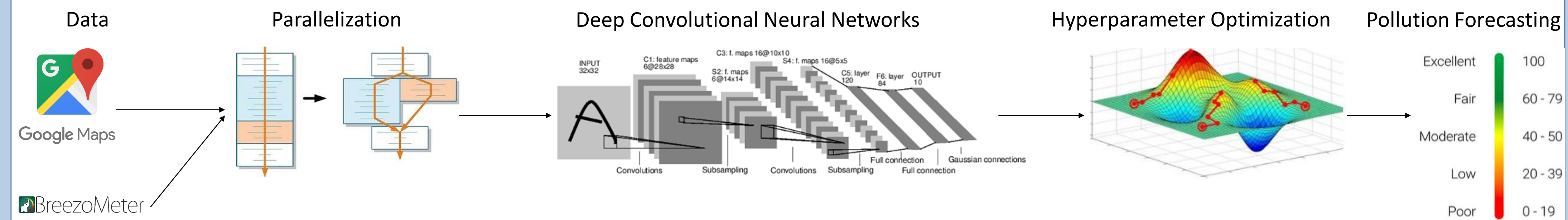
- Advantage: Forecast air pollution with fair accuracy, using data regarding weather, traffic, land use, and satellite sensors
- Disadvantages: Unstandardized, errors in data accumulate to these models

Overall, most existing methods for air pollution forecasting are **unreliable, unscalable, unstandardized, limited, expensive, and/or complex.**

Overview

Attempting to address problems posed by existing methods, this novel study investigates the feasibility and accuracy of using deep convolutional neural networks on satellite images to predict the Breezometer Air Quality Index.

Methodology



Data & Parallelization

The NodeJS package map-dl was used to retrieve satellite images from Google Maps. The Breezometer API was used to fetch air pollution data, including the Breezometer Air Quality Index (BAQI). BAQI, on a scale of 0 to 100, is a holistic air pollution metric that is highly accurate, standardized, and hyperlocal.

First, a list of 57 cities, varying in size, population, location, and air quality, was generated. Google Maps was used to approximate the geographic coordinates of the rectangle that circumscribes each city. Then, mathematical algorithms using trigonometric functions were developed in Java to split each city's rectangle into a grid of smaller rectangles. Next, 7 servers were set up to distribute queries and collect data in a parallelized fashion for each of the 10,000 smaller rectangles across the 57 cities.

An excerpt of the parallelized data collection code developed in Java:

```
List<Server> safeServers = Collections.synchronizedList(servers);
ForkJoinPool fjpool = new ForkJoinPool(safeServers.size());
List<Boolean> success = new ArrayList<Boolean>();
do {
    success = fjpool.submit(() -> {
        return rects.parallelStream().map(rect -> {
            Server server =
                synchronized(safeServers) {
                    server = safeServers.remove(Math.random() * safeServers.size());
                }
            boolean worked = true;
            try {
                // Code for downloading satellite images and air pollution data via
                // Breezometer and map-dl queries
            } catch (Exception e) {
                worked = false;
            }
            synchronized(safeServers) {
                safeServers.add(server);
            }
            return worked;
        }).collect(Collectors.toList());
    }).get();
    for(int i = 0; i < success.size(); i++)
        if(!success.get(i))
            success.remove(i);
            rects.remove(i);
            i--;
    } while(success.contains(false));
}
```

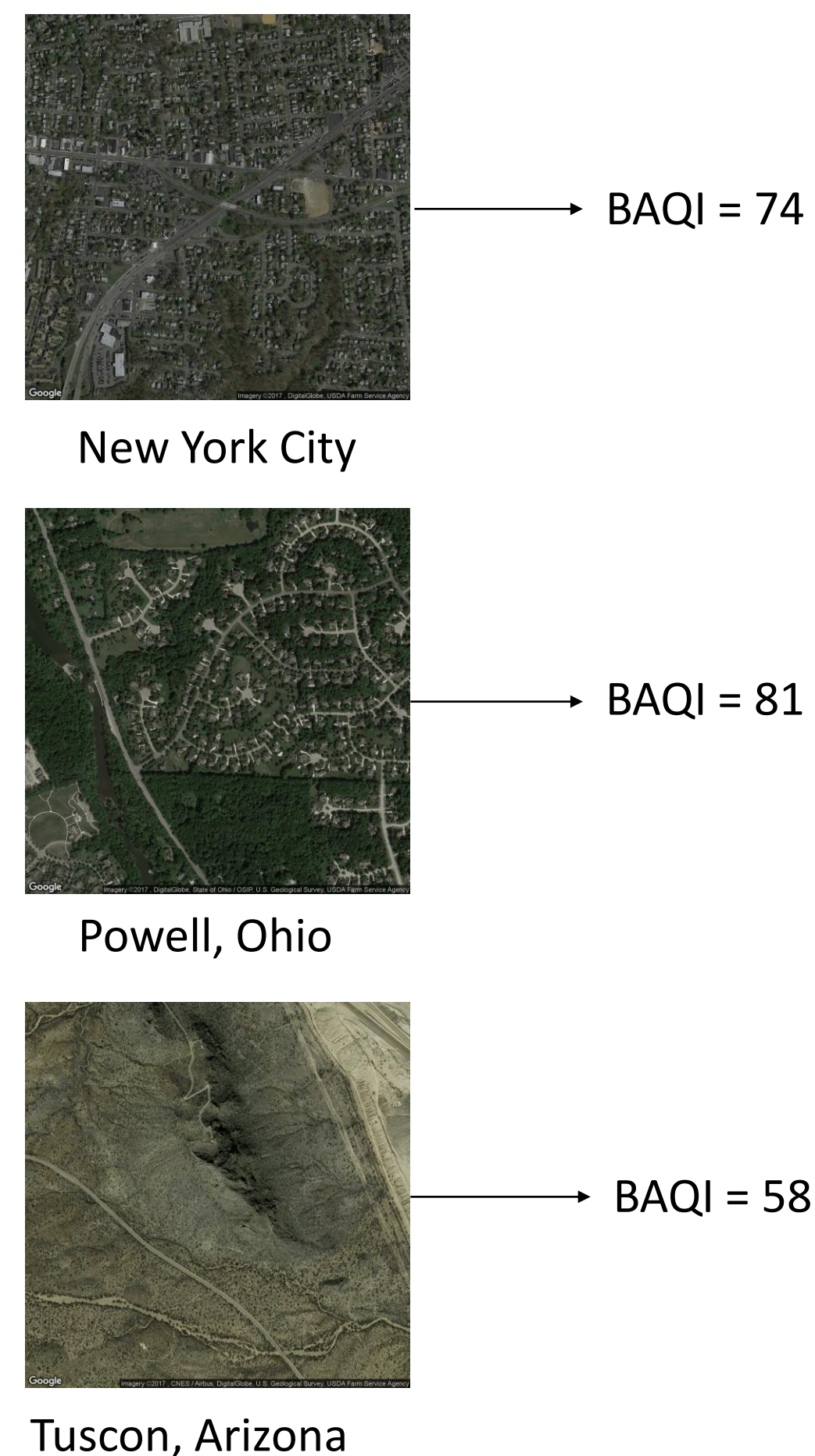
After running the full version of the above code, 10,000 satellite images were downloaded and a CSV of 10,000 air pollution queries was generated. The data collection process took 2 hours.

The 10,000 satellite images were resized from 1280 x 1280 to 200 x 200 using Python's

```
scipy.misc.imresize()
```

An excerpt of the CSV:

1	City	Latitude	Longitude	Date and Time	Air Quality Index (AQI)	Air Quality (AQI) Description	Dominant Pollutant Name	Satellite Image Filename
2	New York City	40.7281	-74.3719	2017-10-31T21:08:32	77	Fair air quality	o3	40.47731035182244-74.36601725124059-40.46831712576325-74.37784487636466-1.0005027310295673.png



Deep Convolutional Neural Networks & Hyperparameter Optimization

An excerpt of the deep convolutional neural networks code using Keras in Python:

```
model = Sequential()
model.add(Convolution2D(nb_filters, nb_conv, nb_conv, activation = 'relu'))
model.add(MaxPooling2D(pool_size=(nb_pool, nb_pool)))
...
model.add(Dropout(0.5))
model.add(Flatten())
model.add(Dense(128, activation = 'relu'))
model.add(Dropout(0.5))
...
model.add(Dense(2, activation = 'softmax'))
```

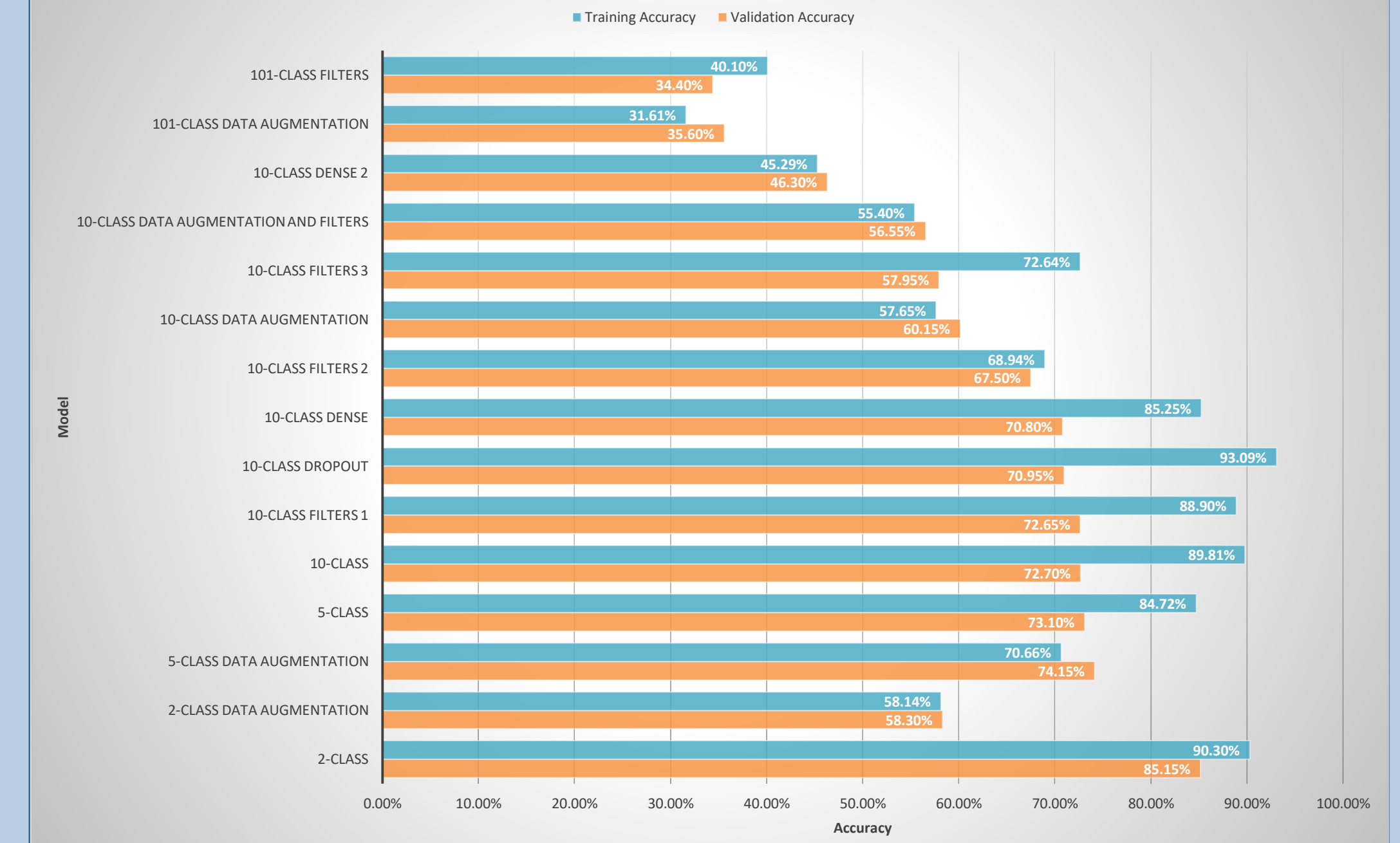
A variety of models were trained on a to predict the BAQI using the satellite images. The following hyperparameters were varied and optimized:

- Number of Convolutional Layers (1, 2, 3, 4, 5, 6)
 - ✓ Allows for prediction of more complex features
 - ✓ Usually increases accuracy but also overfitting
- Number of MaxPooling Layers (1, 2, 3, 4, 5, 6)
 - ✓ Allows for finding features on different scales
 - ✓ Allows for improved efficiency, via reduction of image resolution
 - ✓ Usually increases accuracy
- Size of Convolutional Layers (16, 32, 64, 128, 256, 512)
 - ✓ Increases the number of features that can be detected
 - ✓ Usually increases accuracy but also overfitting
- Dropout (0 – i.e. not used, 0.25, 0.5)
 - ✓ Reduces overfitting by randomly excluding some nodes during training, as this prevents feature detection from relying on any specific nodes
 - ✓ Usually decreases accuracy but increases validation accuracy
- Data Augmentation (Not used, Used)
 - ✓ Allows for finding features that are invariant to transformations
 - ✓ Easier and more memory efficient than manual generation of more data
 - ✓ Usually increases accuracy for object detection problems
- Output Loss Estimators
 - ✓ 2-class (binary) cross-entropy
 - ✓ 10-category cross-entropy
 - Number of categories (intervals of 0 to 100 scale) can give a sense of the model's accuracy
 - The model may also consider close or almost-correct values as wrong, especially as the number of categories approaches the data variance

The models were trained on a Titan X Pascal GPU in roughly half an hour each.

Results & Discussion

Model vs. Training and Validation Accuracies



- For most of the models, the training and validation accuracies were close in value, indicating the models were fit well. Some of the 10-class models overfit, with larger training accuracies than validation accuracies.
- As the number of categories increased, the validation accuracy decreased.
- The effect of data augmentation in the models was unclear, as the 5-class and 101-class models performed better with data augmentation, but the 2-class and 10-class models performed worse with it.
- Models with convolutional layers of size 16, 32, and 64 performed better than those with 32, 64, and 128, which performed better than those with 16, 16, and 16.
- Models with dropout values of 0.5 performed better than those with 0.25.
- Models with 3 dense layers performed better than those with 2, which performed better than those with 1.

Conclusions & Key Advantages

Using deep convolutional neural networks on satellite images was not only feasible but also highly accurate in predicting the Breezometer Air Quality Index. Specifically, the models achieved accuracies of 85.15% on 2-class (binary) classification of air pollution, 74.15% on 5-class classification, and 72.70% on 10-class classification. This novel research is the first to the best of my knowledge to predict air pollution levels from satellite images. There are several key advantages of this research and its methods:

1. **Reliable** – Satellite images are highly accurate and always available.
2. **Scalable** – Satellite images are accurate at a variety of scales.
3. **Standardized** – Satellite images and BAQI are standardized.
4. **Not limited** – Satellite images are unlimited in geographic coverage.
5. **Inexpensive** – Satellite images are free to use.
6. **Simple** – Satellite images are digital and simple to use.