AIR POLLUTION CHARACTERIZATION AND PREDICTION USING DEEP AUTOENCODER MODEL

Nick Shannon[1], Raul Chavez[2], Vandana Nunna [3]

1. Loyola University Chicago – Department of Computer Science, 2. Department of Electrical and Computer Engineering at the University of Texas – Rio Grande Valley, 3. University of North Texas - College of Engineering

COLLEGE OF ENGINEERING
Department of Computer Science & Engineering

The University of Texas Rio Grande Valley

LOYOLA UNIVERSITY CHICAGO
Preparing people to lead extraordinary lives

MOTIVATION

- Air pollutants are both a cause and a symptom of climate change. In order to more effectively combat climate change, it is important to understand and characterize the source of emissions. We developed a technique to characterize cities by the levels of various air toxins using an undercomplete autoencoder model.

- We compare two techniques for reducing the input dimensionality - principle component analysis (PCA) and undercomplete autoencoders.

AUTOENCODER

- An autoencoder is a neural network that encodes its input data into a low-dimensional latent space encoding. Once the input is encoded, it can be decoded by reconstructing the input using the latent space.
- Representing pollution data in a lower dimensional space can improve performance on tasks such as classification, and can make data visualization easier. One particularly useful application is clustering cities based on one or two of the couple of the most relevant. dimensions

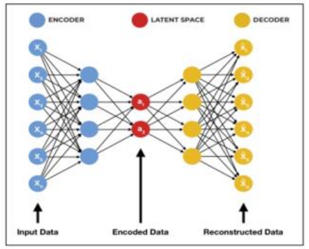
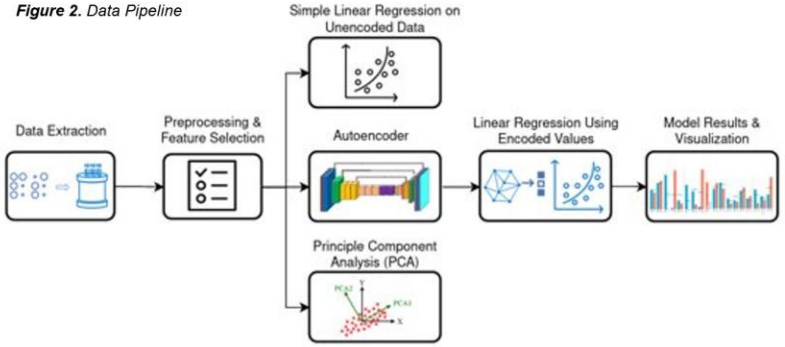


Figure 1. Autoencoder Architecture

METHODOLOGY

Figure 2. *Data Pipeline*


Data Extraction
Preprocessing & Feature Selection
Simple Linear Regression on Unencoded Data
Autoencoder
Principle Component Analysis (PCA)
PCA2
PCA1
Y
X
Linear Regression Using Encoded Values
Model Results & Visualization

We created a dataset of pollution levels in numerous cities throughout the United States using OpenWeather's Air Pollution API. We preprocessed the incoming data by removing null values and duplicate entries.

OpenWeather Pollution API	
Time period	6.5 months
Time frequency	Daily
# of cities	18,526
# of features	120 (originally 190)
# of component gases	8
Dependent variable	Single day of time series (3/27/21)

Figure 3.
Dataset Characteristics

* Component gases include CO, NO, NO2, O3, SO2, NH3, PM2.5, and PM10.

- A simple linear regression was performed on the encoded data that was generated by both the PCA and AE models. Additionally, we performed a linear regression on the unencoded values for comparison.
- Each model's goodness of fit was evaluated using the percentage of explained variance for the three linear regressions.

Optimal Hyperparameters for Carbon Monoxide (CO)	
Activation function	Leaky ReLU
Loss function	Mean Squared Error
Optimizer	Adam
Learning rate	0.001
Batch size	64
Epochs	150

Figure 4.
Model Tuning

- A grid search was conducted to find the best parameter by testing a cartesian product of various learning rates, batch sizes, and epochs.
- Leaky ReLU gave us the best results of the activation functions we tried.

RESULTS

We compared the results from principal component analysis (PCA) and the undercomplete autoencoder models. We found that the PCA discovered lower dimensional linear representation whereas the autoencoder was capable of learning non-linear relationship between gases. Higher percentage explained variance of Ozone (O3) represents that its concentration in the environment does not vary as much as that of particulate matter (PM.5).

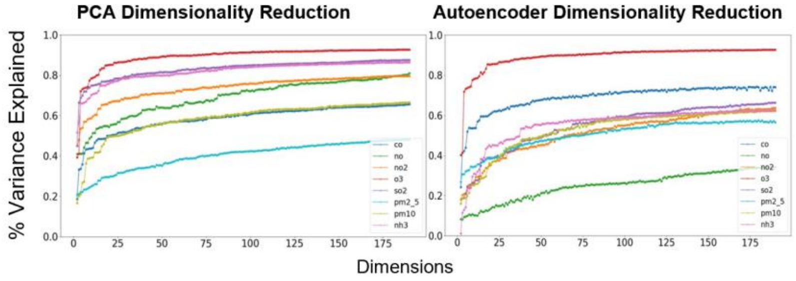


Figure 5. *Graph of PCA (Left) and Autoencoder (Right) reduced representation of Polluting Gases*

We saw a large jump in explained variance within the first two dimensions for both the PCAs and the autoencoders. We clustered the cities based on the first two latent dimensions of carbon monoxide in order to make inferences on how the model characterizes each city.

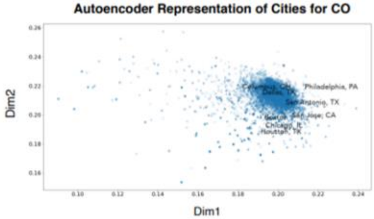


Figure 6. *Plot with labels of highly populated cities in first two dimensions*

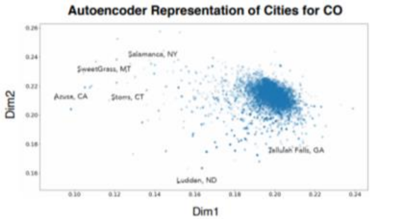


Figure 7. *Plot with labels of outlier cities in first two dimensions*

CONCLUSION

The embeddings produced by the autoencoder can be used to identify various patterns and qualities of air pollution. The embedding could be used to analyze NH3 emissions from industrial farming and animal husbandry, tropospheric O3 emissions from internal combustion engines and power plants, and NO2 levels caused by automobiles. These are just a few of the numerous use cases that could help us better understand air pollution.