Deep Learning Based Sensitivity of Numerical Model

Zongrun Li
Georgia Institute of Technology
Atlanta, GA
zli867@gatech.edu

Abstract

In this project, we train two different convolutional neural network models based on three years numerical air quality model (CMAQ) inputs and outputs. To evaluate the impacts from different inputs, we use the trained model to calculate gradient of regression results and define two different metrics to reflect impacts from each factors based on Jocobian matrix. By comparing the structures of different CNN models and different gradient definitions, we rethink application of CNN models in scientific area.

1. Introduction

Deep learning model has been proved that it has good performance on capturing complex non-linear relationship prediction. Raissi et al. has conducted neural network to solve partial differential equations (PDEs) including hyperbolic, advection-diffusion and Naiver-Stokes which are widely used in atmosphere research [5]. CMAQ which simulates physics and chemistry based on numerically solving PDEs is used for air quality research. Since solving numerical model is computationally intensive, researchers have tried to use numerical model results to train machine learning or deep learning based empirical model which is more computationally effective. Sayeed et al. conducted a naive CNN to predict time series of ozone values and the result shows a good performance of pollutants prediction [6]. Xing et al. combines emission reduced scenario based on DDM (decoupled direct method in three dimensions) [3] results and CMAQ results to evaluate different emissions impacts on air pollution [8]. In this work, we also want to evaluate the impacts on air pollution from different factors such as emissions or meteorology. Unlike Xing et al. [3], we use gradient to evaluate the impact of different factors. The gradient calculation is based on previous work from Simonyan et al. [7] while unlike calculating the gradient of classifier, we focus on saliency map of regression in deep learning.

2. Approach

2.1. Data and Feature Selection

In this work, we select CMAQ input data and output results provided by EPA [1]. We extract three years data including 2013, 2015 and 2017 in California, United States. Since the limitations on time and computational resources, we only includes features from meteorology, anthropogenic emissions on surface level as model input. We use daily maximum 8-hour ozone as training label. The input features include 32 meteorology features and 23 emission features which covers two major ozone precursors volatile organic carbon (VOC) and $\mathrm{NO_x}$ related chemical species (NO, $\mathrm{NO_2}$, HONO). Notice that the CMAQ is a 3D grid based numerical model. Since we only focus on surface level, each features has 2D structure (115×65 for covering CA). And the data is withheld 10% as test data set.

2.2. Structure of Deep Learning Models

The first model is based on the work from Sayeed et al.[6]. We use AlexNet as our CNN structure [2]. We treat each feature as a layer in image processing. Unlike RGB three layers in images, we have 55 layers for each input. We use single input channel CNN to predict all grid values (a vector of ozone) in CMAQ output. MSE is used for loss function to do the regression training.

The second model is a multi-channel CNN model. Since from previous understanding [4], ozone has non-linear relationship with VOC and $\mathrm{NO_x}$ emissions. The effect from these two emissions do not have any correlations with each other. Also, meteorology does not have strong correlation with anthropogenic emissions. We decide to separate all features into three categories VOC, $\mathrm{NO_x}$ and meteorology and each category will be an independent input channel for the multi-channel CNN. We apply 4 Conv-Batch Norm-ReLU-Maxpool for meteorology and VOC and 3 for $\mathrm{NO_x}$ since $\mathrm{NO_x}$ includes less features.

2.3. Gradient of Regression Model

The regression models we applied here can be described as a projection from vector $I \in \mathbf{R^m}$ to vector $O \in \mathbf{R^n}$. Notice that the multi-channel CNN is also a vector to vector projection if we consider the input as the concatenate of all channels. So, we will have a Jacobian matrix when we evaluate $\frac{\partial O}{\partial I}$. To evaluate the impacts from all grids in input domain to a specific grid i in output domain, we define prediction-wised gradient as:

$$gradient_k = \sum_{i=1}^{m} \frac{\partial y_k}{\partial x_i}$$

This gradient tells us the what factors affect the ozone in a specific region. If we want to decrease ozone in a specific region, we could know the direction by the prediction-wised gradient.

Similarly, we can define another source-wised gradient which is used to evaluate the impacts from specific domain of emission source or metereology condition to the whole domain ozone:

$$gradient_k = \sum_{i=1}^{n} \frac{\partial y_i}{\partial x_k}$$

This could tell us which region have positive or negative impacts on the ozone in whole CA. It can be used to identify high-pollution sources.

3. Experiments and Results

We use Adam as the optimizer and train each model 50 epochs. Notice that 90% of data are used for training and 10% of data are used for testing.

For single channel CNN, the MSE between prediction and true values are less than 25 after 50 epochs for both test and training data. For multi-channel CNN, the MSE between prediction and true values are less than 20 after 50 epochs for both test and training data (Figure 1). To evaluate the model performance, we use the trained deep learning model to predict time series of ozone concentration in CA and yearly averaged ozone distribution (Figure 2, 3). The temporal pattern has been captured while multi-channel CNN has better performance than single channel CNN. For spatial distribution, the prediction has high correlation with CMAQ output while the maximum difference is about 5 ppb for both two models. It can be expected from MSE values and the results are consistent with previous research [6, 8].

To evaluate the sensitivity of each factors, we calculate two different gradients based on previous definition. To understand whether the meteorology and emissions have positive impact or negative impact on ozone, we normalize the gradient for each species (layers) and sum the normalized

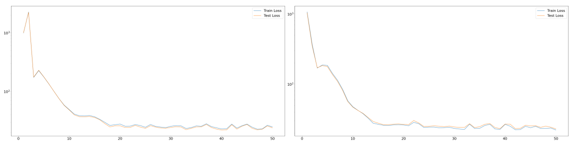

Figure 1. MSE loss versus training epochs number. Left hand side is single channel CNN and right hand side is multi-channel CNN

gradient by these three categories: VOC, $\mathrm{NO_x}$ and meteorology. For source-wised gradient (figure 4), two different CNN models show similar patterns for VOC impacts. Big cities such as Los Angeles has high VOC sources which has relatively important impact on CA ozone. For $\mathrm{NO_x}$, both of models show positive values at the center of some big cities while single channel CNN shows negative impacts near the positive points. For meteorology, these two models show different gradient patterns. One thing need to pay attention is that the results from single channel CNN has high correlation among these three categories. Also for prediction-wised gradient, the gradient results for three different categories from single channel CNN also have high correlation among each other. This result is not realistic since meteorology should not be expected strong related with emissions. It can be explained by the covolutional calculation process. Convolutional calculation may have good performance when we know different layers have correlation such as the RGB in images while this correlation could not be correct when we apply the convolution for numerical model. While for multi-channel CNN, these three categories do not have strong relationship with each other since these convolutional processes are independent in these three channels. A favourable pattern can be observed from the prediction-wised result in multi-channel model is that the positive impacts of $\mathrm{NO_x}$ has correlation with major roads in CA which implies the main contribution of ozone in these region is from transportation.

4. Conclusion

This work is a preliminary exploration on building deep learning based air quality model. We successfully reproduce the temporal and spatial patterns of ozone based on meteorology and emission information. We define two different gradients in regression model for sensitivity evaluation. While the structure of deep learning model should be improved in the future work. Although feature selection is not required for deep learning models, we need to pay more attentions on some potential assumption in the structure definition. For example, convolution is a calculation applied to each layers simultaneously. Rotate or shift image could be captured in this process. While rotation or shifting could convey really different information when we apply convolution to numerical models. Also, the gradient should be sta-

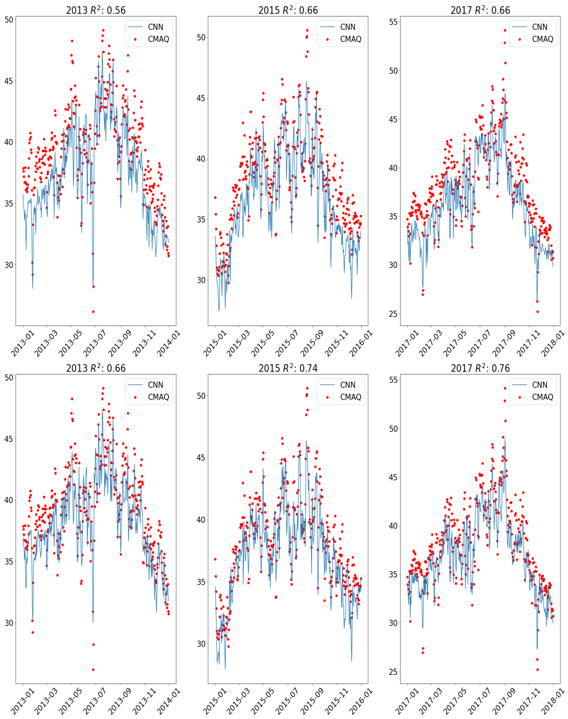

Figure 2. Time series prediction in selected three years. Upper three figures are results from single channel CNN and lower figures are multi-channel CNN results.

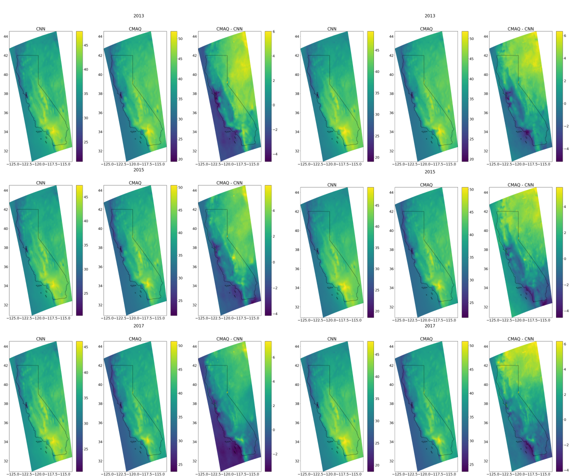

Figure 3. Yearly spatial distribution. Left hand side is from single channel CNN and right hand side is from multi-channel CNN

ble when we have enough data from finite difference view. In this work, differences between gradient results from two models could be partly explained by the size of data (about 1000 days data). For future work, we should include more CMAQ results and embed understanding of features in designing multi-channel CNN model. Also, other structure such as LSTM should be explored.

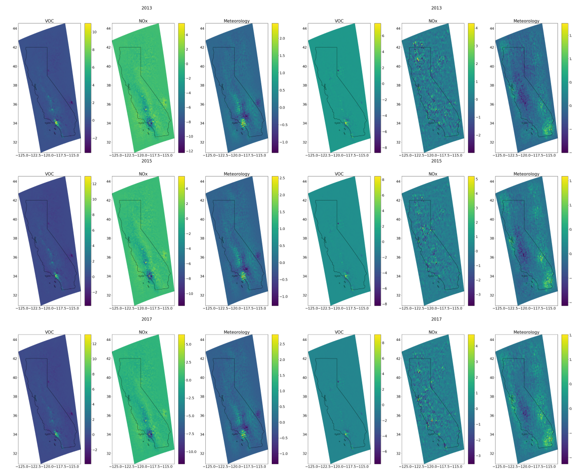

Figure 4. Source-wised gradient. Left hand side is from single channel CNN and right hand side is from multi-channel CNN

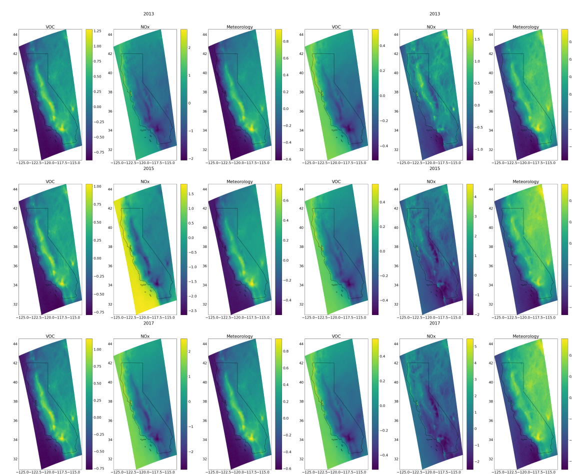

Figure 5. Prediction-wised gradient. Left hand side is from single channel CNN and right hand side is from multi-channel CNN

References

[1] US EPA. EQUATESv1.0: Emissions, WRF/MCIP, CMAQv5.3.2 Data – 2002-2017 US$_12kmandNHEMI_108km$, 2021. 1

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 1

[3] Sergey L Napelenok, Daniel S Cohan, Yongtao Hu, and Armistead G Russell. Decoupled direct 3d sensitivity analysis for particulate matter (ddm-3d/pm). *Atmospheric Environment*, 40(32):6112–6121, 2006. 1

[4] Yu Qian, Lucas RF Henneman, James A Mulholland, and Armistead G Russell. Empirical development of ozone isopleths: Applications to los angeles. *Environmental Science & Technology Letters*, 6(5):294–299, 2019. 1

[5] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations. *arXiv preprint arXiv:1711.10561*, 2017. 1

[6] Alqamah Sayeed, Yunsoo Choi, Ebrahim Eslami, Jia Jung, Yannic Lops, Ahmed Khan Salman, Jae-Bum Lee, Hyun-Ju Park, and Min-Hyeok Choi. A novel cmaq-cnn hybrid model to forecast hourly surface-ozone concentrations 14 days in advance. *Scientific reports*, 11(1):1–8, 2021. 1, 2

[7] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 1

[8] Jia Xing, Shuxin Zheng, Dian Ding, James T Kelly, Shuxiao Wang, Siwei Li, Tao Qin, Mingyuan Ma, Zhaoxin Dong, Carey Jang, et al. Deep learning for prediction of the air quality response to emission changes. *Environmental science & technology*, 54(14):8589–8600, 2020. 1, 2