Predicting Air Quality with Deep Learning

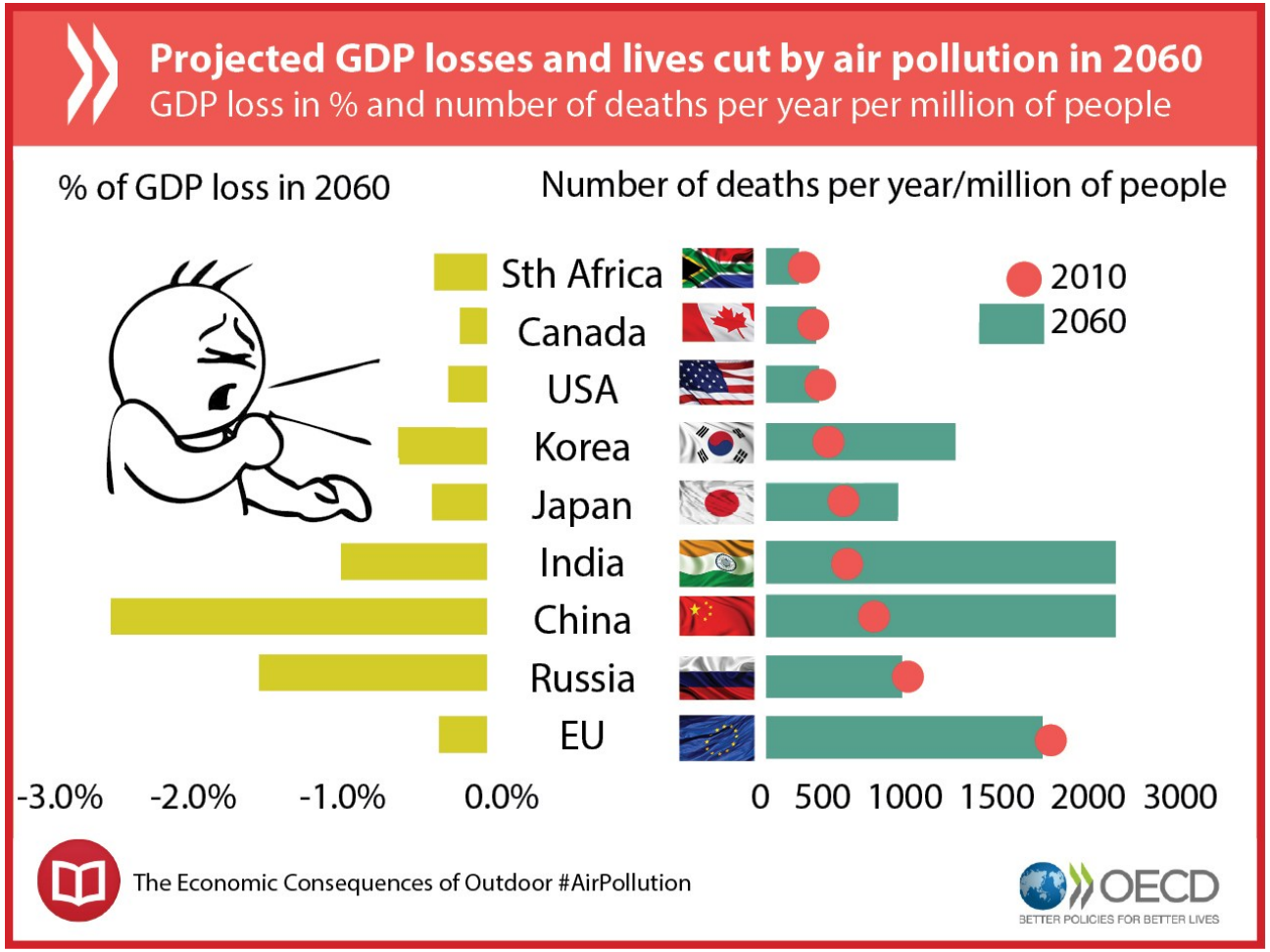
CS325b final presentation

March 11, 2018

Adele Kuzmiakova
Emanuel Cortes
Max Evans

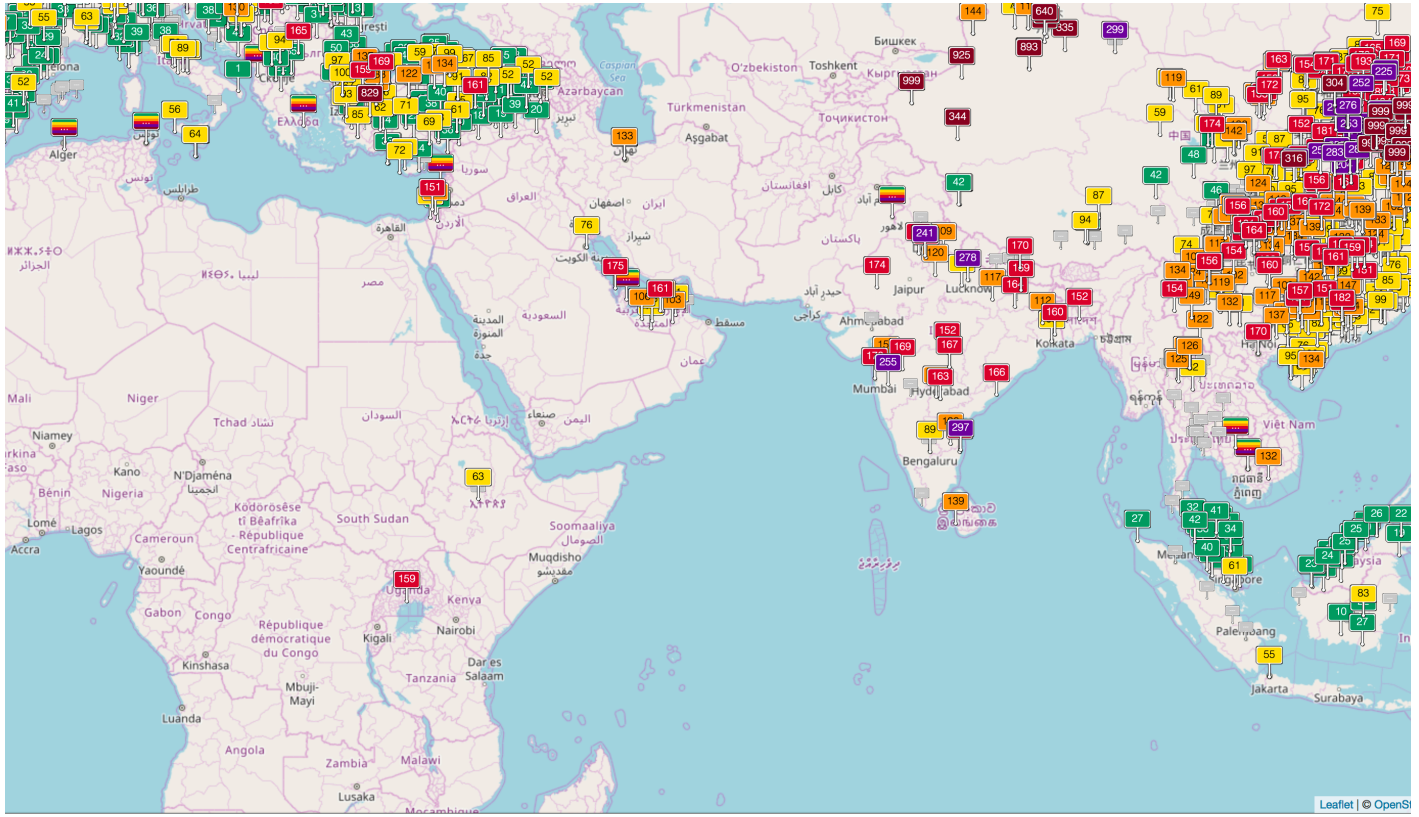
Why Air Quality?

- Air pollution is a global problem
- **India and China:** death toll will be > 2 million of people/year by 2060
- Affects economic activity: GDP loss up to 3%



Problems with Measuring Air Quality

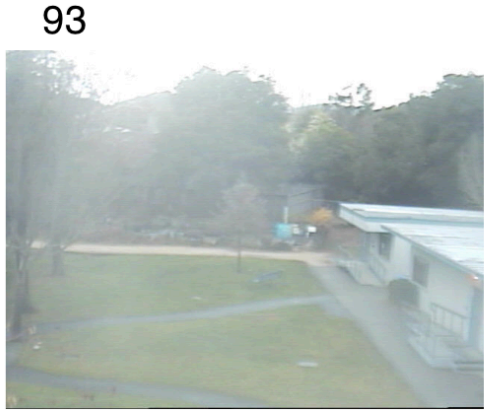
- **Data scarcity:** very few monitoring stations in India, almost none in Africa
- **Expensive:** costs for one station are higher than $10,000
- But this would really help decision-makers

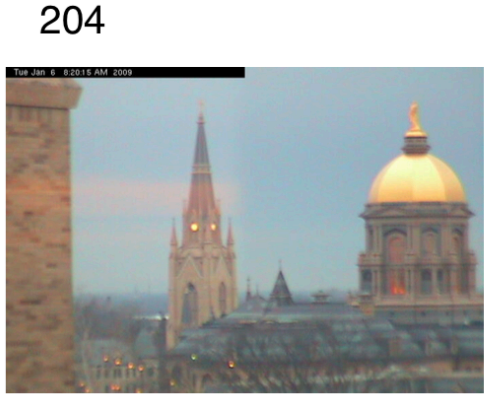


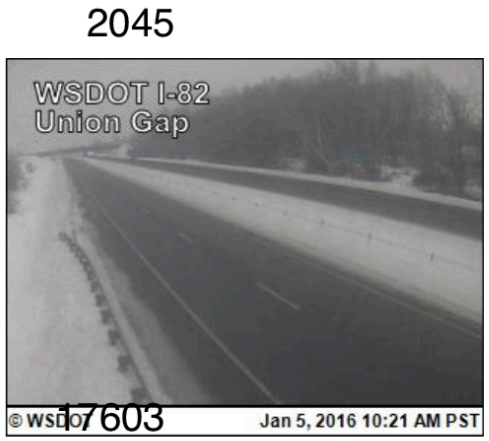
Webcam Images

- Ubiquitous, easy to obtain, cheap
- Long time-series, global-scale
- Social media

Can we infer outdoor air quality from the time-series of public images?





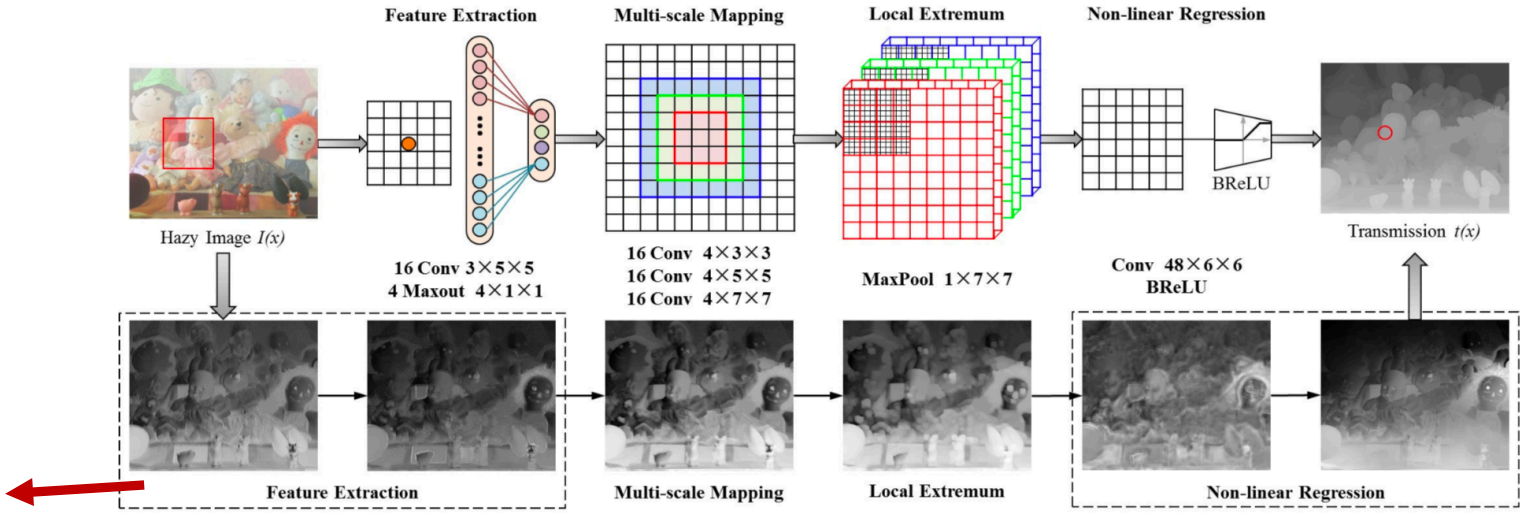


Previous Work

- Single image haze removal:

Haze-relevant features:
1) Dark channel
2) Atmospheric light
..

transmission



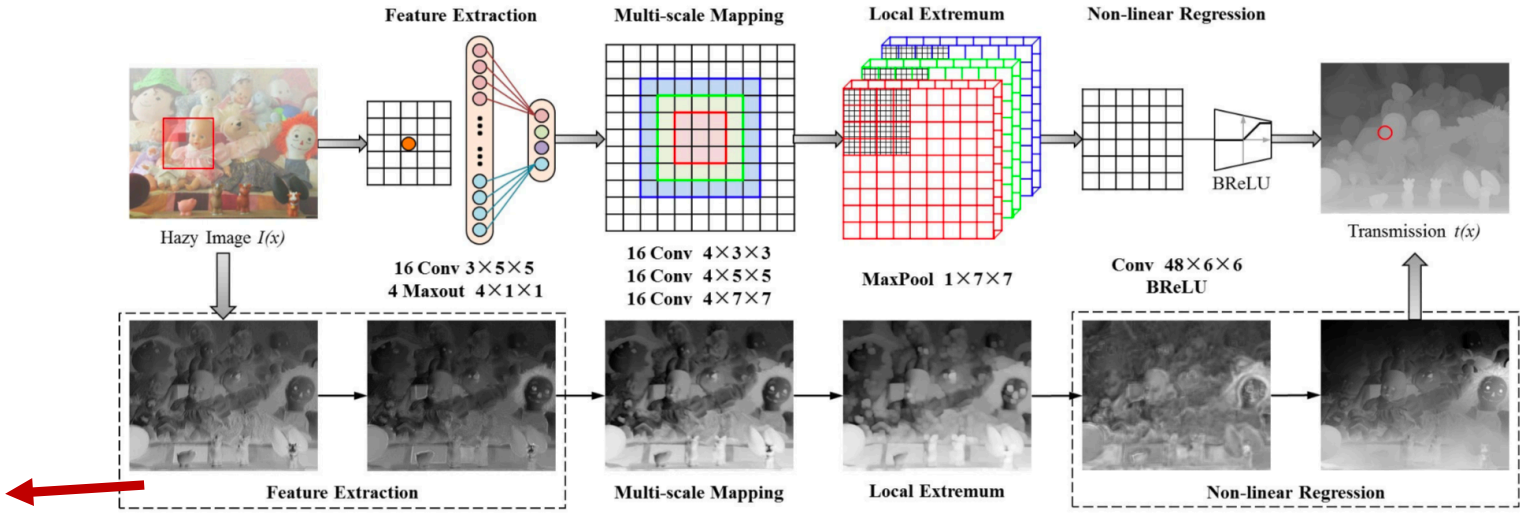
Cai et al: DehazeNet: An End-to-End System for Single Image Haze Removal

Previous Work

- Single image haze removal:

Haze-relevant features:
1) Dark channel
2) Atmospheric light
..

transmission



Cai et al: DehazeNet: An End-to-End System for Single Image Haze Removal

- Using deep learning for webcam images:

Table 2: Air quality index for PM2.5 and PM10.

Level	1	2	3	4	5	6
PM2.5 Range	<35	35-75	75-115	115-150	150-250	>250
PM10 Range	<50	50-150	150-250	250-350	350-420	>420

Table 4: Coincidence matrix for PM2.5 by different methods on

	1	2	3	4	5	6
1	326	13	3	1	3	2
2	63	48	11	1	0	4
3	33	23	23	8	4	1
4	11	10	4	5	8	3
5	15	4	2	10	43	9
6	3	2	2	4	12	44

(PAPLE)

	1	2	3	4	5	6
1	305	25	14	0	4	0
2	57	47	19	1	2	1
3	25	25	30	6	6	0
4	7	8	7	5	9	5
5	8	3	22	5	30	15
6	8	0	4	0	18	37

(CNN-Softmax)

	1	2	3	4	5	6
1	296	44	0	0	8	0
2	53	68	0	0	5	1
3	28	56	0	0	7	1
4	4	22	0	0	13	2
5	6	24	0	0	46	7
6	5	5	0	0	28	29

(LReLU-CNN-Softmax)

	1	2	3	4	5	6
1	293	35	3	6	6	5
2	61	45	9	1	2	9
3	30	25	14	10	6	7
4	11	8	4	6	9	3
5	15	5	0	11	38	14
6	3	1	2	1	15	45

(ReLU-CNN-Softmax)

Zhang et al:
On Estimating Air Pollution from Photos Using Convolutional Neural Network

Main Questions

1) What features may be the most relevant for inferring air quality from the images?

2) Do time-series of public images contain enough information for deep learning methods to predict the air quality?

Haze Features

1) Transmission
2) Dark channel
3) Atmospheric light
4) Saturation
5) Power spectrum

6) Local features: hour, day, month

7) Weather features

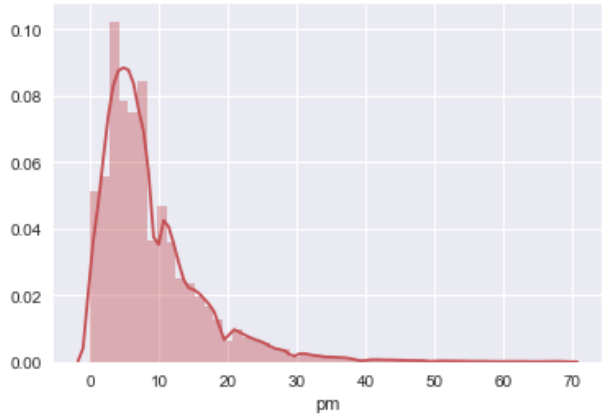
Original

Transmission

ElasticNet for feature selection:

$$\min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda \left[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1 \right]$$

Data



Public images of roads and skylines

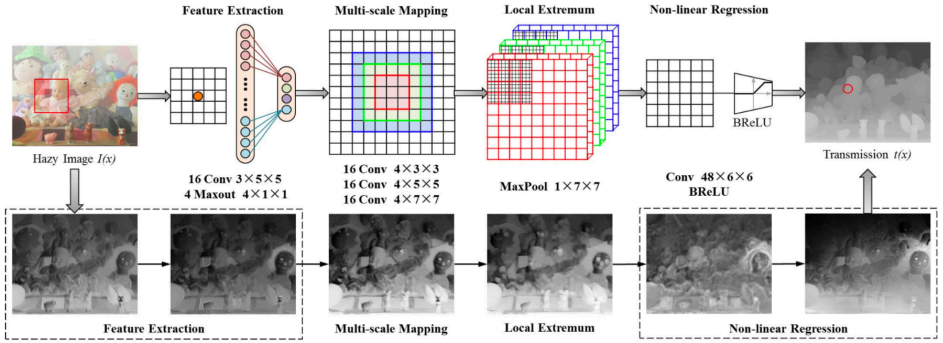
800+ locations

pm label distribution

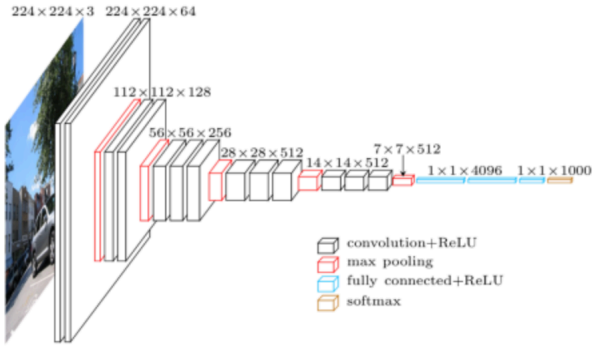
- So far we focused on the site in Alaska

Deep Learning Architectures

- DeHazeNet

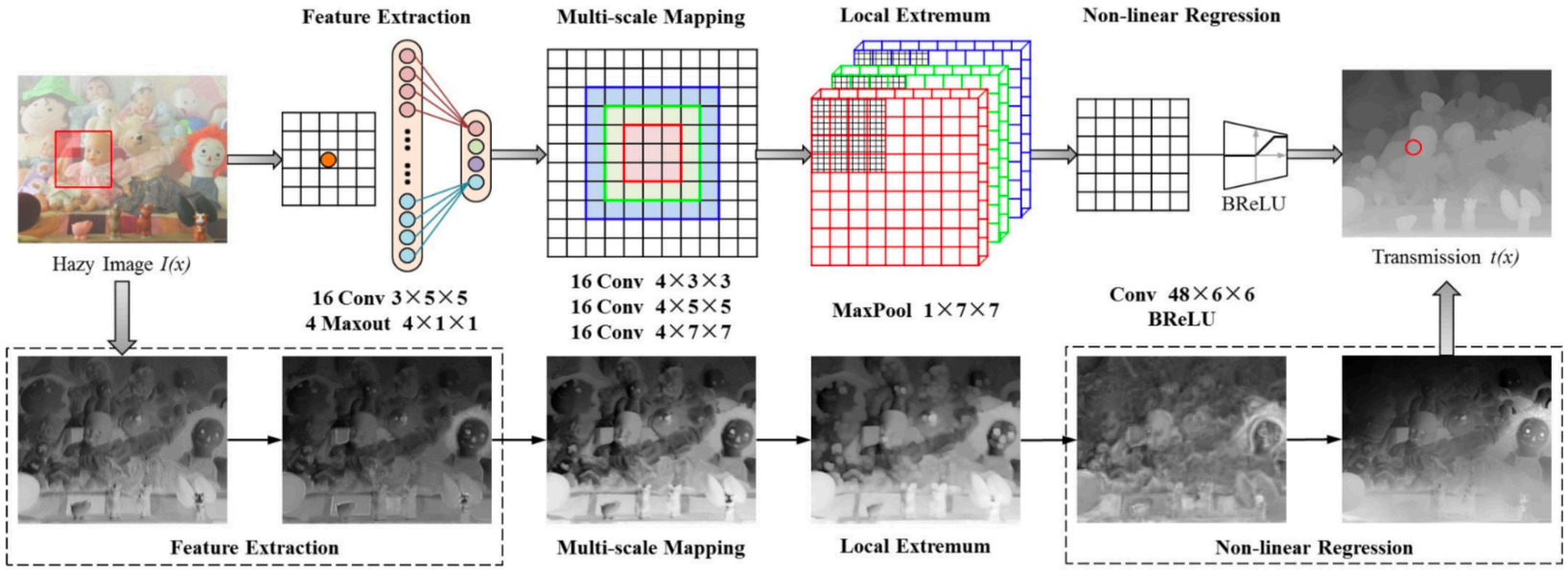


- VGG 16

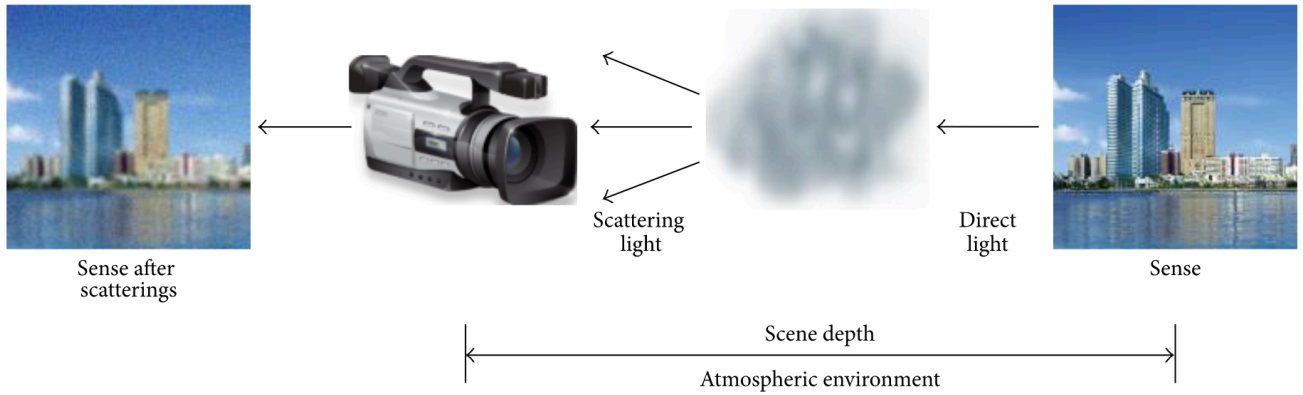


Deep Learning

- Original DeHazeNet

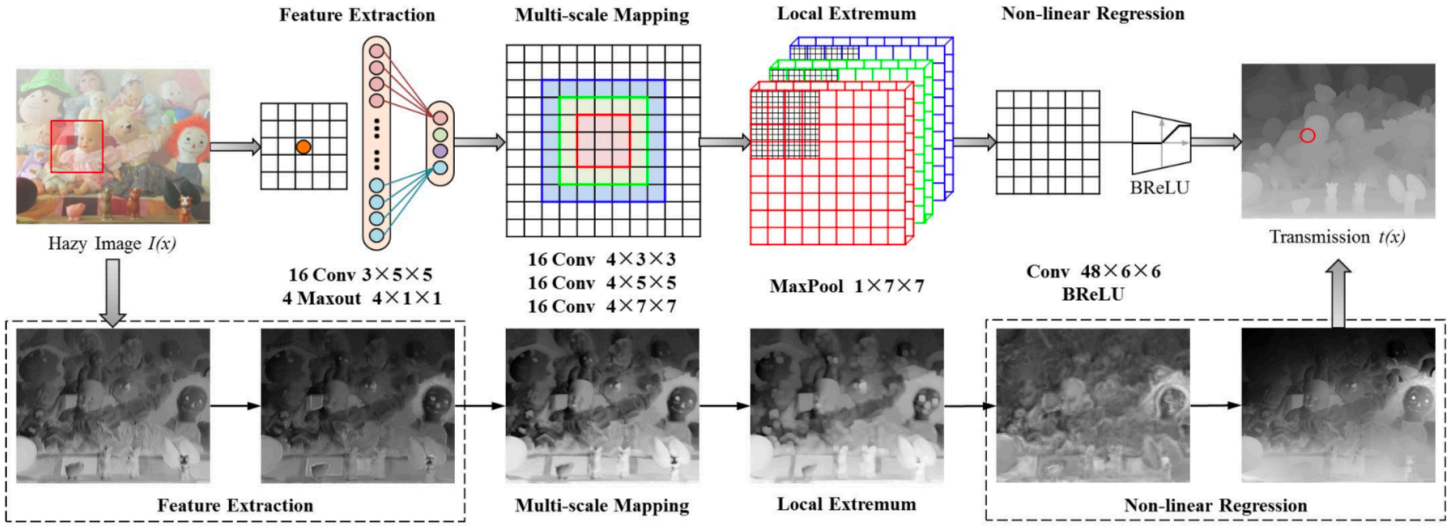


- What is it trying to model?

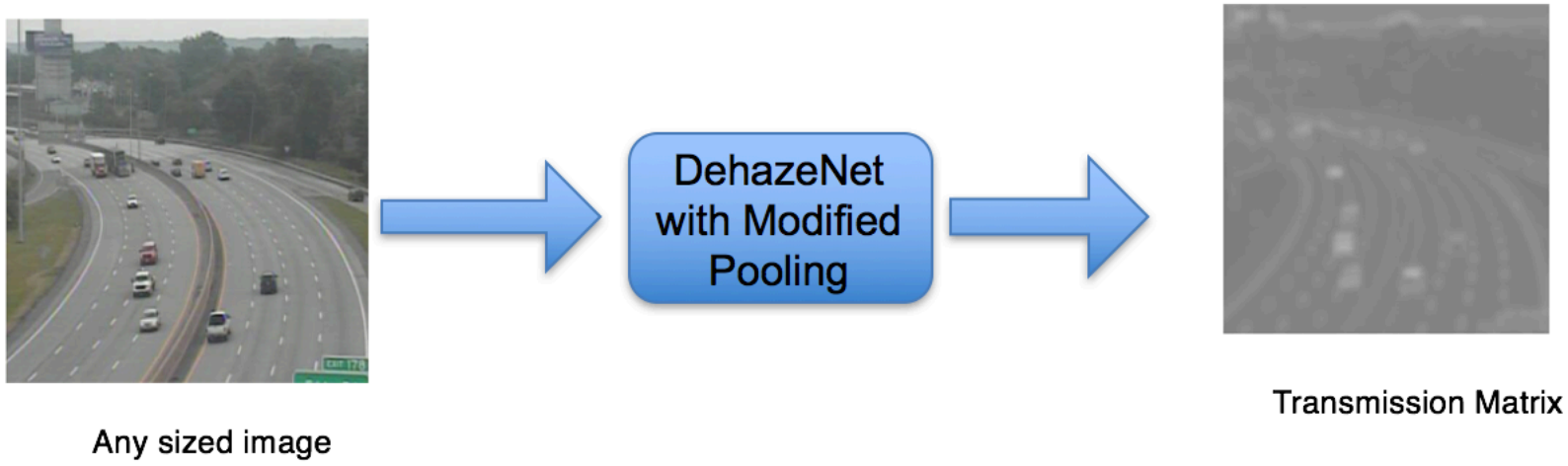


Deep Learning

- Original DeHazeNet

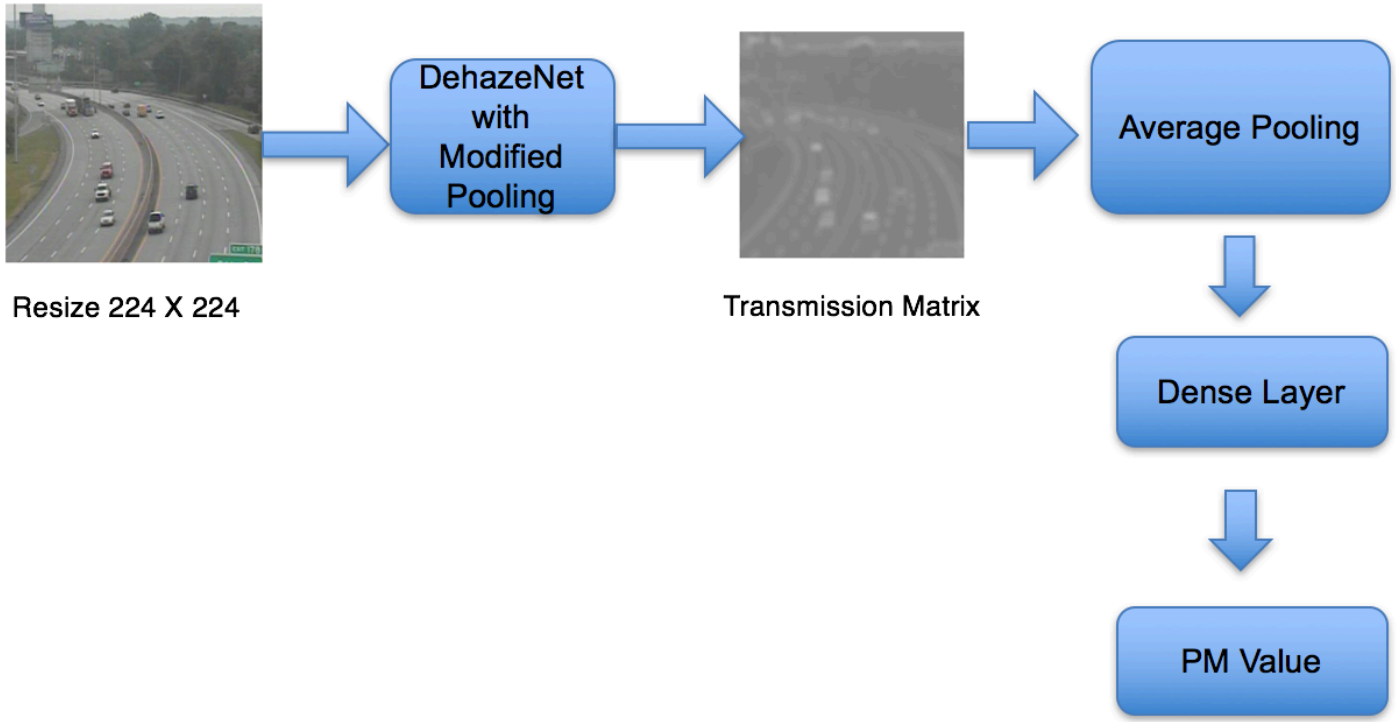


- Modified DehazeNet



Deep Learning

- Modified DehazeNet with PM Value

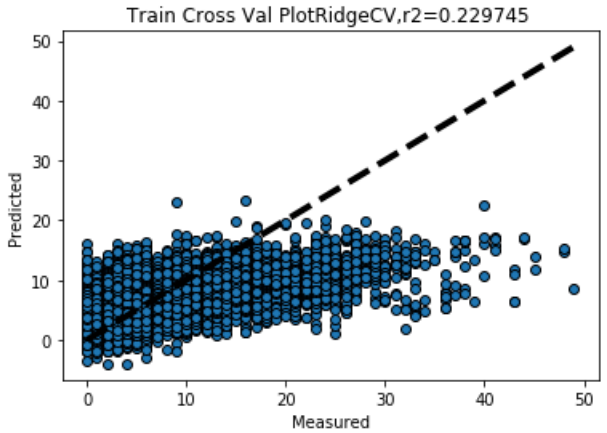


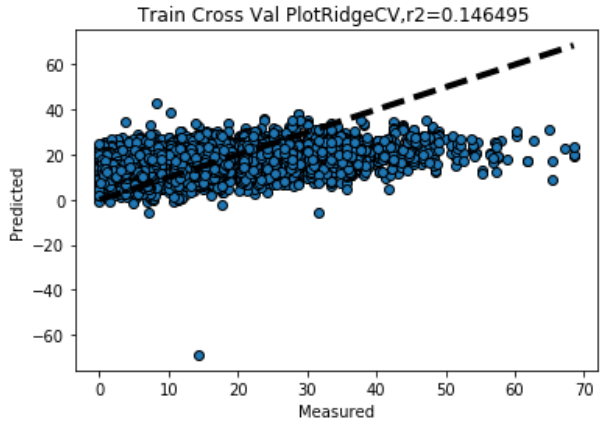
Results

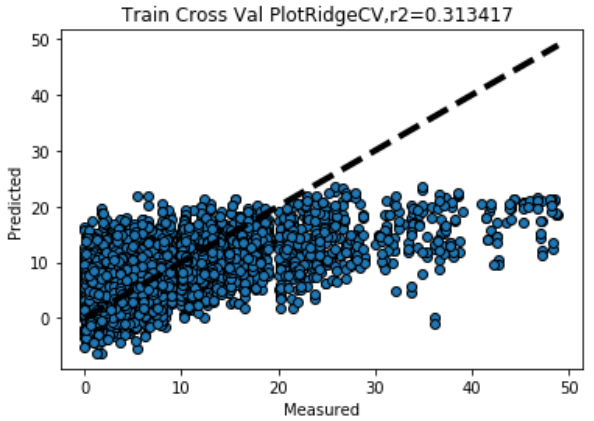
- Ridge Regression
- Elastic Net
- Deep Learning Architectures
- VGG 16

Results – Ridge Regression

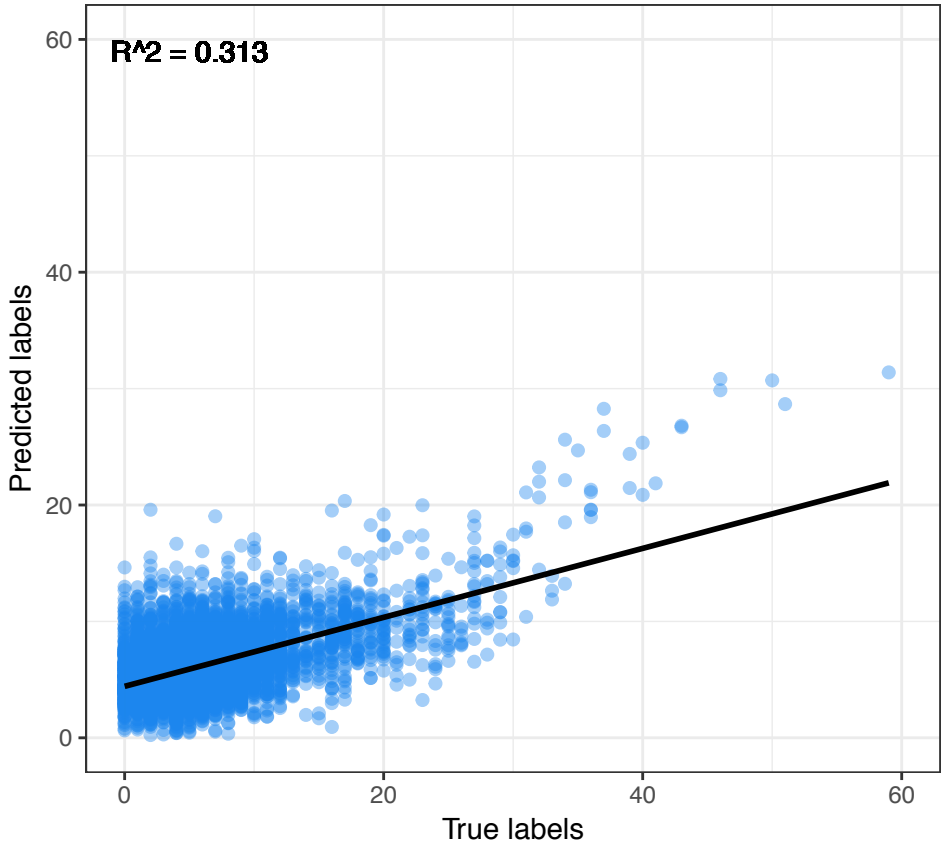
Webcam	Split	Model	Features	HyperParam Alpha	Train	Val
18879	Random Shuffle	Ridge	haze+weather+date	0.001	0.45	0.43
18879	Time Split	Ridge	haze+weather+date	0.1	0.42	0.28
18879	Time Split	Ridge	weather+date	0.1	0.24	0.12
18879	Time Split	Ridge	haze	0.1	0.31	0.20
1066	Time Split	Ridge	haze+weather+date	0.1	0.33	0.11
17603	Time Split	Ridge	haze+weather+date	0.1	0.28	-0.04
21587	Time Split	Ridge	haze+weather+date	0.1	0.32	0.05





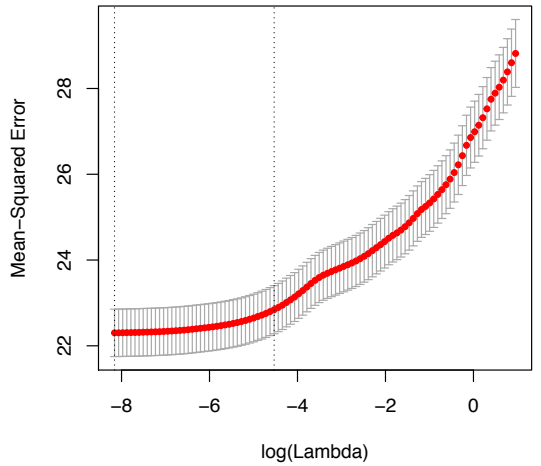


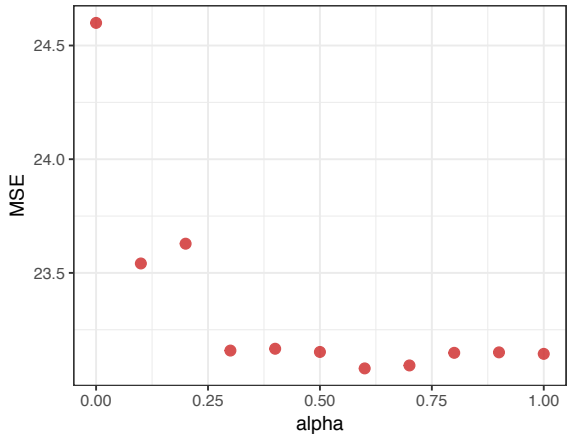
Results - ElasticNet



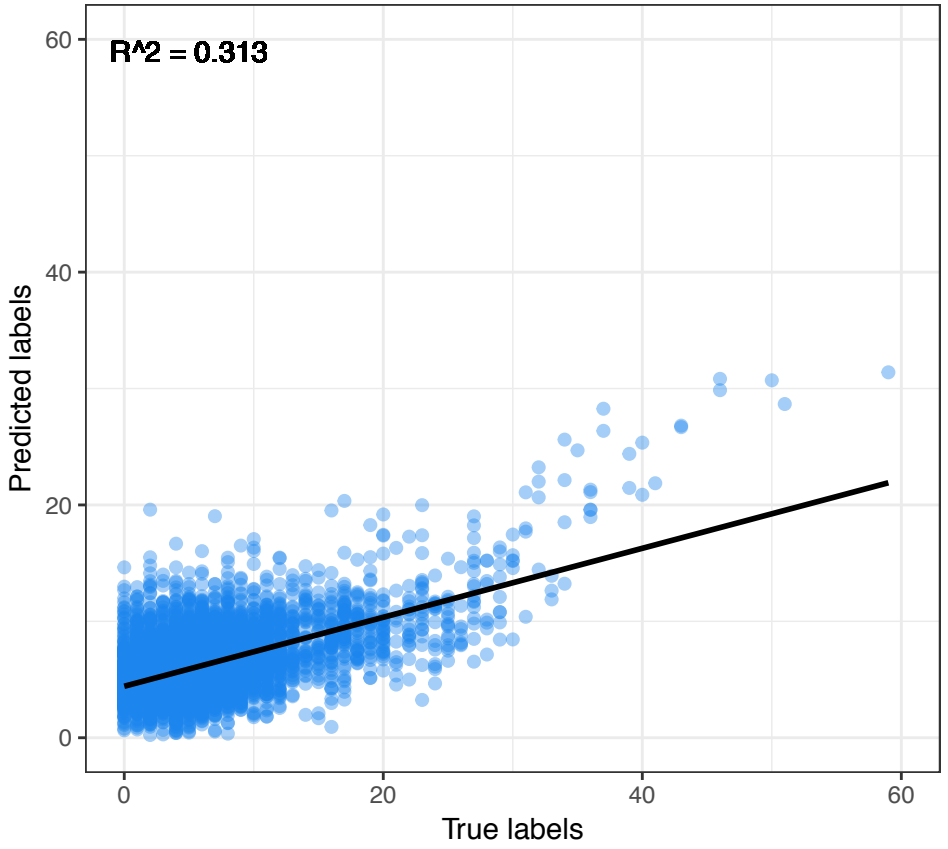
Possible bias due to sunlight:
True = 0, predicted = 15

Finding lambda* and alpha* on the validation set:

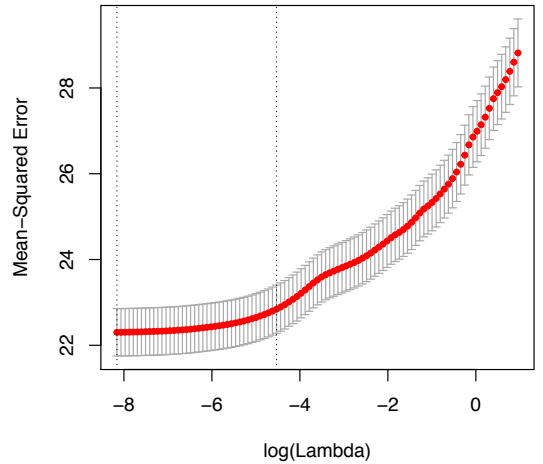


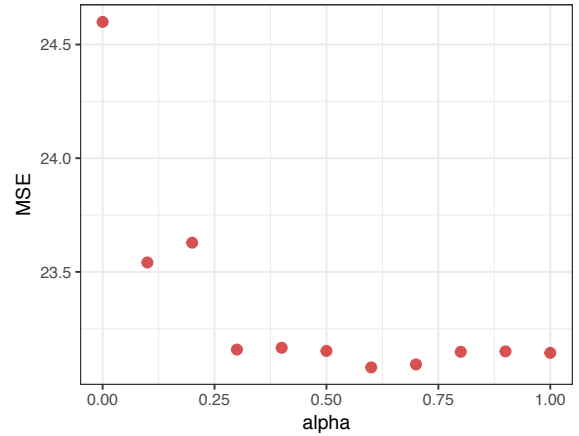


Results - ElasticNet



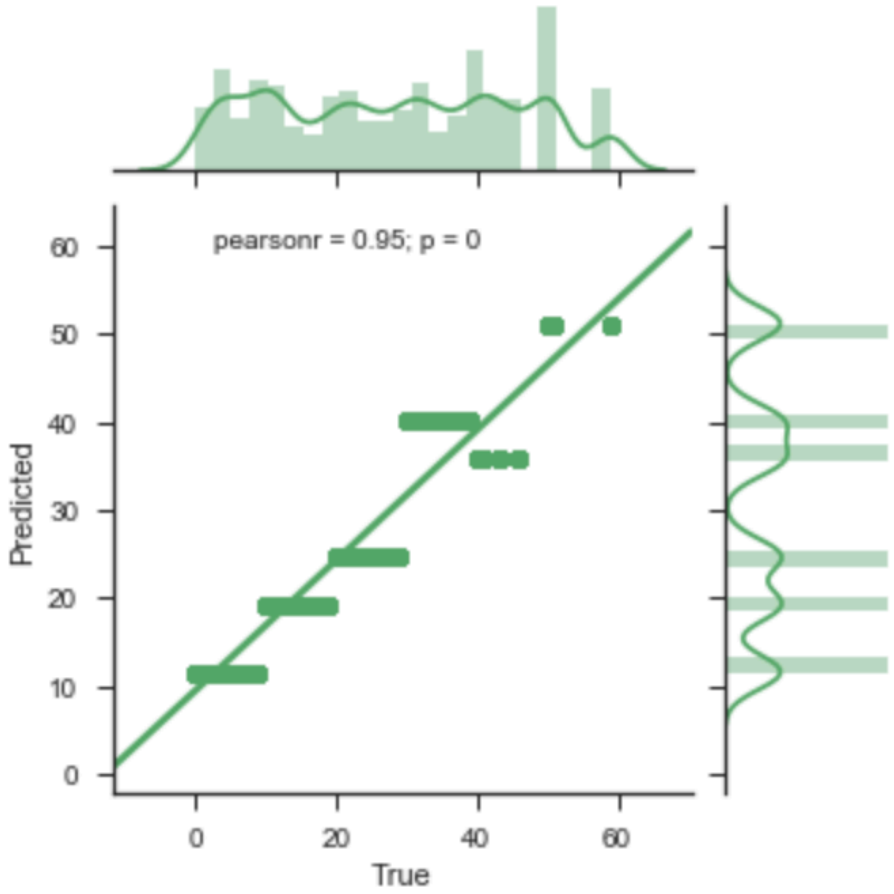
Finding lambda* and alpha* on the validation set:





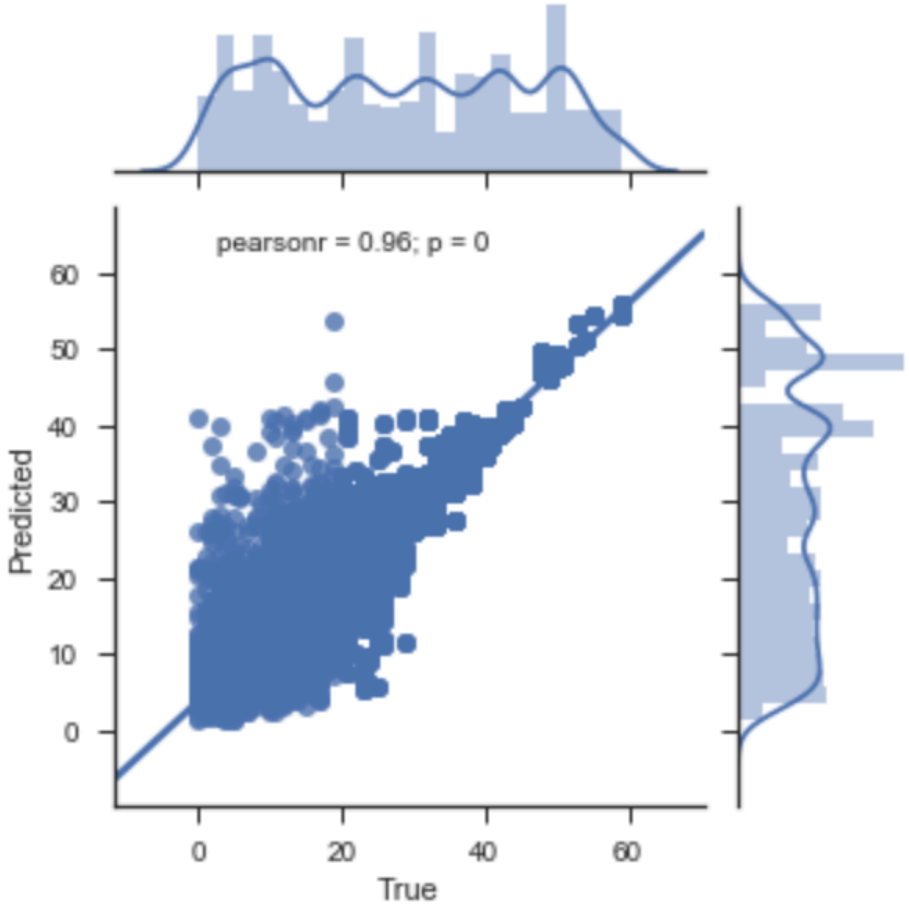
Results – VGG: Webcam 1066

VALIDATION



R^2: 0.86

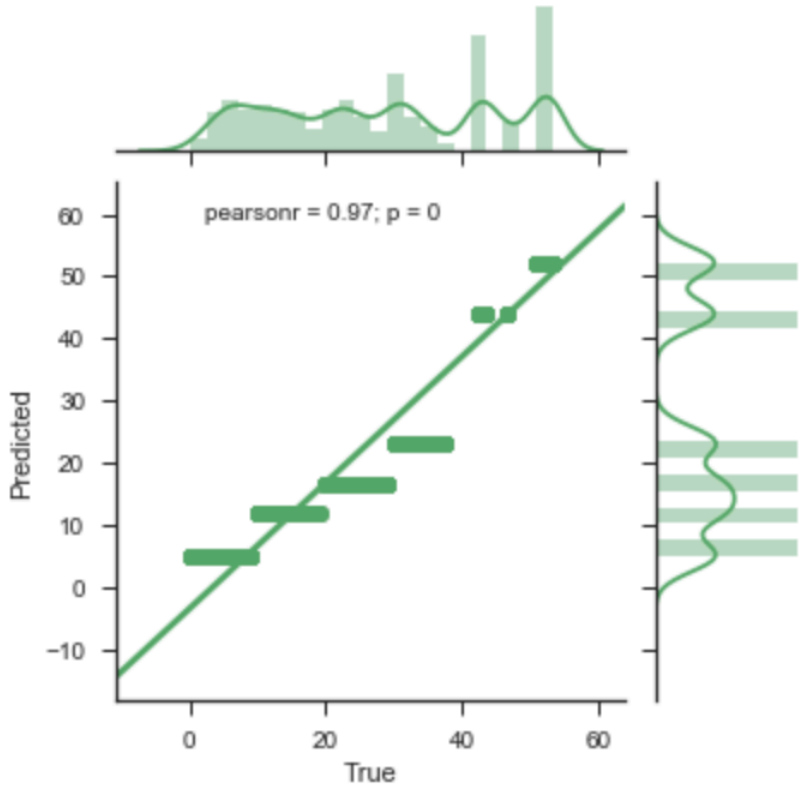
TRAINING



R^2: 0.92

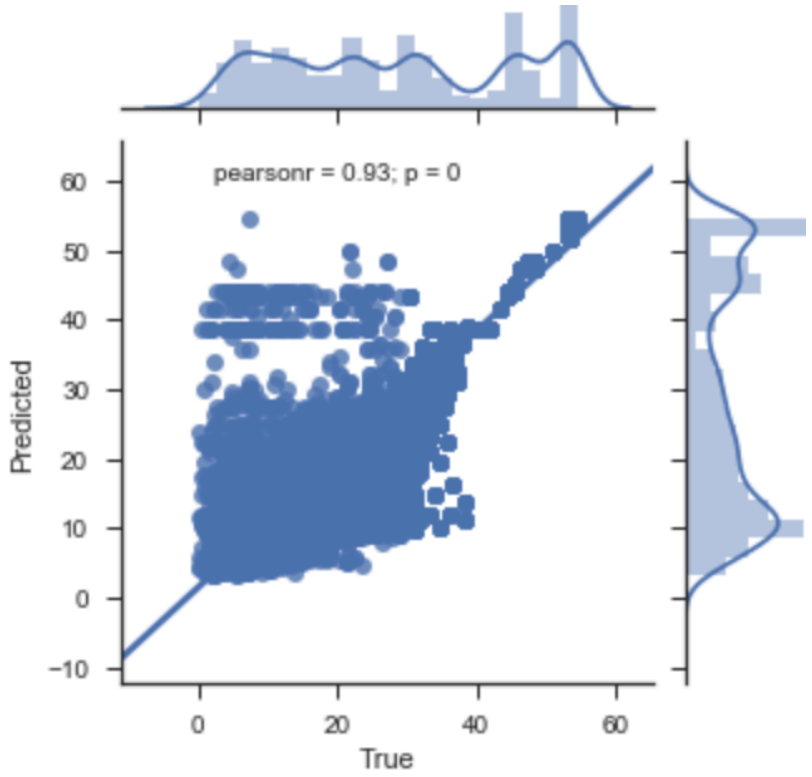
Results – VGG: Webcam 17603

VALIDATION



R^2: 0.90

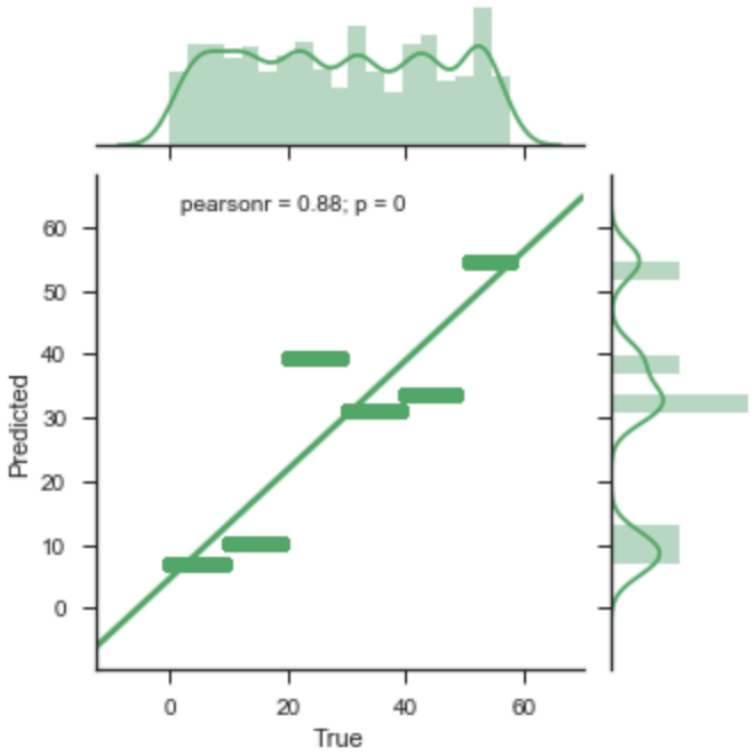
TRAINING



R^2: 0.85

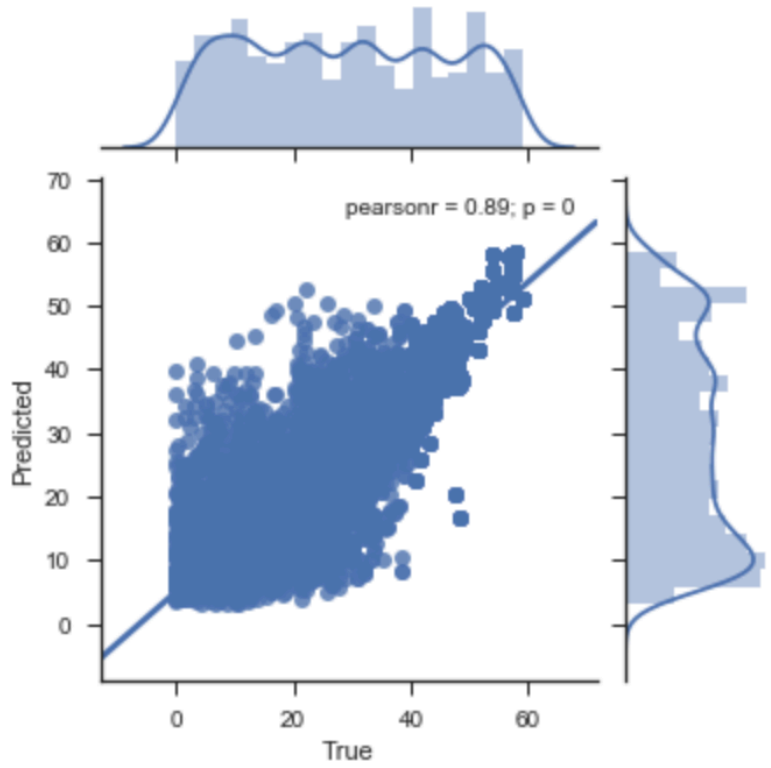
Results – VGG: Webcam 21587

VALIDATION



R^2: 0.76

TRAINING



R^2: 0.80

Results – Summary

Webcam	Split	Model	Features	Train	Val
18879	Random Shuffle	Ridge	haze+weather+date	0.45	0.43
18879	Time Split	Ridge	haze+weather+date	0.42	0.28
18879	Time Split	Ridge	weather+date	0.24	0.12
18879	Time Split	Ridge	haze	0.31	0.20
1066	Time Split	Ridge	haze+weather+date	0.33	0.11
17603	Time Split	Ridge	haze+weather+date	0.28	-0.04
21587	Time Split	Ridge	haze+weather+date	0.32	0.05
1066	Random Shuffle	VGG	raw image	0.92	0.86
1066	Random Shuffle	DehazeNet	raw image	0.54	0.4
17603	Random Shuffle	VGG	raw image	0.85	0.9
17603	Random Shuffle	DehazeNet	raw image	0.63	0.47
21587	Random Shuffle	VGG	raw image	0.8	0.76

Conclusions

- There is predictive power on images
  1) Haze features superior to metadata based features

- Haze features are able to predict PM value
  1) Additional data can be included

- Currently overfitting on specific location
  1) Useful for webcams as PM monitors
  2) More generalization needed for non time-series images

Future Work

- Generalization of our model
  1) Across Time
  2) Across Space

- ResNet using batch normalization avoid overfitting

- Use non time-series images, open air images

- Modeling the concentration of other pollutants from haze features