DHS SCIENCE AND TECHNOLOGY

DeepXplore: Automated Whitebox Testing for Neural Networks

Homeland Security

Science and Technology

Barry Masters, Transportation Security Laboratory

John Tatarowicz, Battelle

Brett Brillhart, Battelle

October 17, 2018

Science and Technology Directorate

So What? Who Cares?

- Space: DeepXplore can be used for testing Deep Learning (DL) based Automatic Target Recognition (ATR) algorithms in Advanced Imaging Technology (AIT) systems.
- Problem: The blackbox nature of neural networks can make it difficult to identify learned features and edge case examples
- Solution: DeepXplore's Automated Whitebox Testing Framework
- Conclusion: Utilized DeepXplore to create image augmentations realistic to Advanced Imaging Technology (AIT) systems and test ATR algorithms.
- Future Work:
    - Refine image augmentations to cover realistic bounds of change and extend AIT augmentations to cover adversarial augmentations.
    - Design physical data collection to match synthetically generated data and quantify weaknesses in algorithm performance.

DeepXplore Testing

- Uses unlabeled test inputs to generate new, synthetic inputs using augmentations that both activate a large number of neurons within a DNN and cause similar DNN's to behave differently.

- Paper: DeepXplore – Automated Whitebox Testing of Deep Learning Systems https://arxiv.org/abs/1705.06640

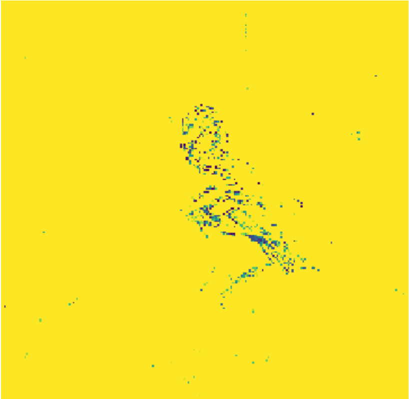
- Github: https://github.com/peikexin9/deepxplore

DeepXplore with ImageNet

Example from DeepXplore runs with ImageNet

Orig: All Brambling

Light:
VGG16: Ruffled Grouse
VGG19: Brambling
ResNet50: Brambling



Lighting difference invisible to human eye caused one model to misclassify

DeepXplore with AIT Algorithms

- Created image augmentations realistic to Advanced Imaging Technology (AIT) systems to test ATR algorithms.

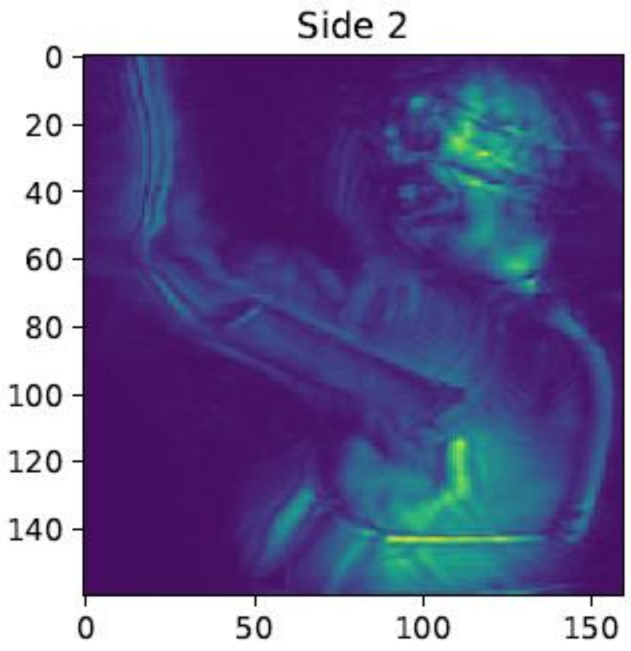
- Blurs to simulate moving arms, horizontal bars to simulate dead sensors.

- Added data collection features such as heatmaps and scatter plots.

Image Augmentations: Lighting

Original Image

Image: 20161014_0001 Zone: 3
Prediction: 0.9992055 Truth: 1
Iteration: 0



Transformation: LIGHT Behavior: Incorrect

Image: 20161014_0001 Zone: 3
Prediction: 0.0059422944 Truth: 1
Iteration: 1

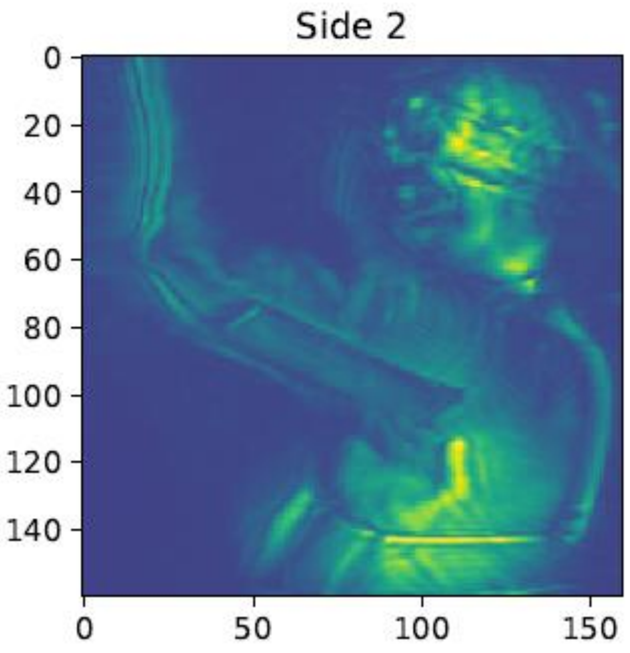


Image Augmentations: Dead Detector

False Negative

Transformation: BAR Behavior: Incorrect

Image: 20161014_0002 Zone: 15
Prediction: 0.27435794 Truth: 1
Iteration: 7

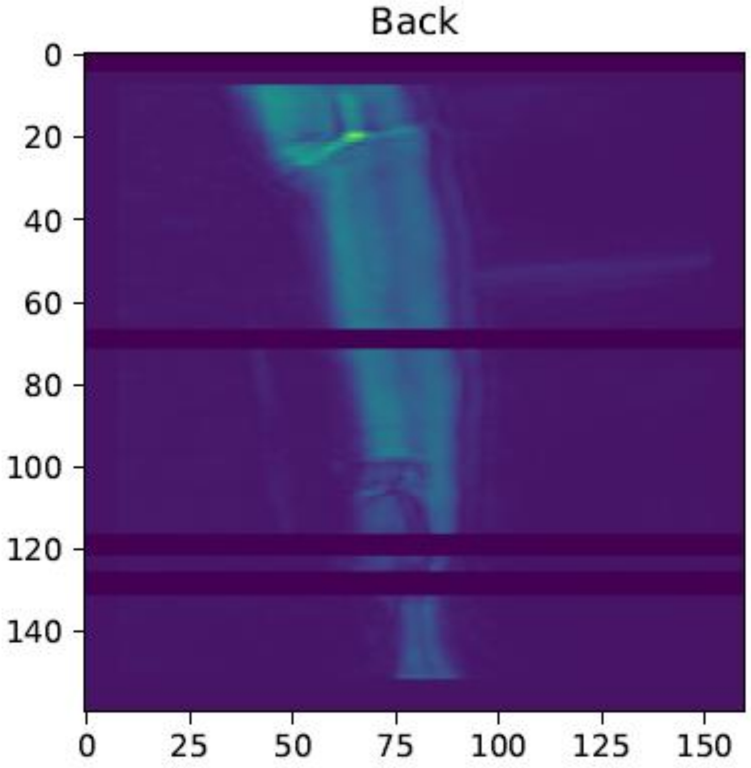
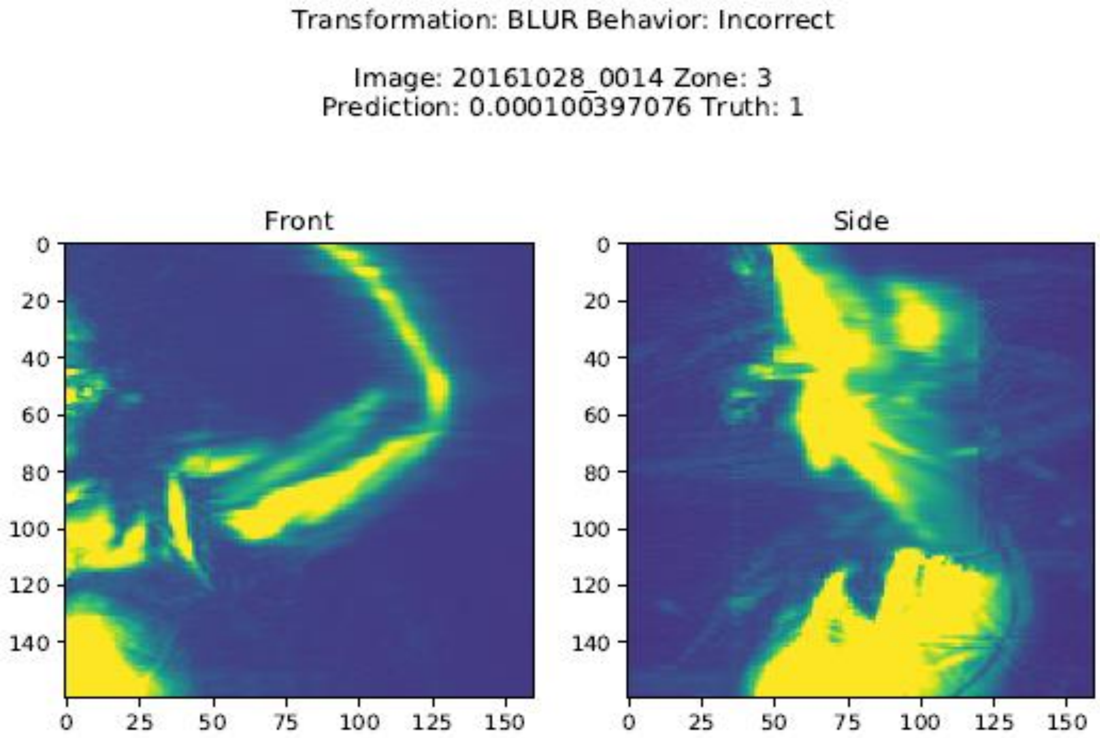


Image Augmentations: Blurs

False Negative

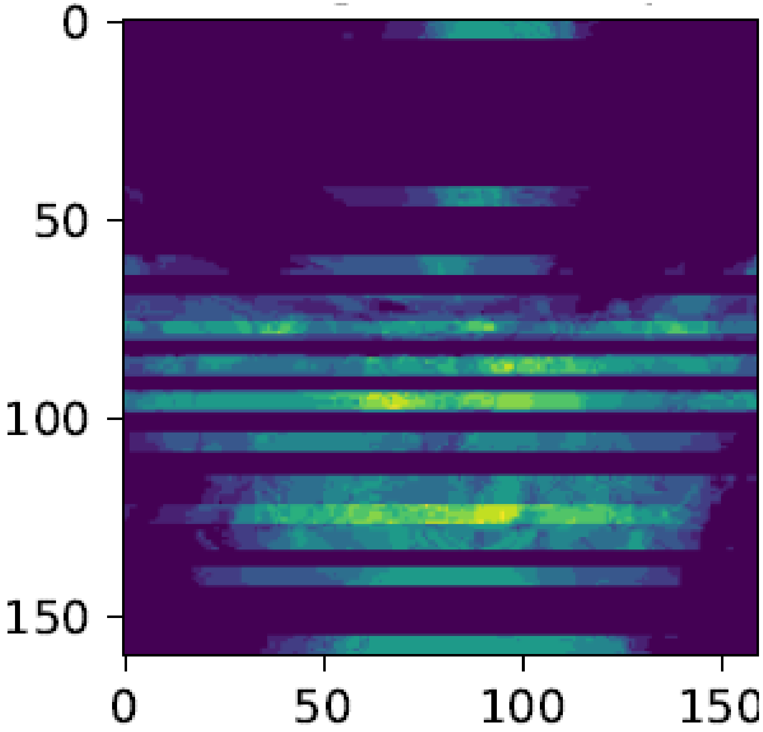


Transformation: BLUR Behavior: Incorrect

Image: 20161028_0014 Zone: 3
Prediction: 0.000100397076 Truth: 1

Data Collection: Heatmaps

Zone 5 Heatmap



Future Plans for DeepXplore

- Integration with other test algorithms.
- Refine system specific image augmentations to cover realistic bounds of change.
- Extend AIT augmentations to cover adversarial augmentations.
- Design physical data collection to match synthetically generated data.
- Analyze and quantify weaknesses in test algorithm detection performance.
- Extend to another detection modality (CT, projection X-ray).

Point of Contact(s)

Barry Masters

AIT DT&E Technology Lead

Transportation Security Laboratory

Barry.Masters@hq.dhs.gov

(609) 813-2722

Brett Brillhart

Junior Technician

Battelle

brillhart@battelle.org

(989) 615-4390

John Tatarowicz

Research Scientist

Battelle

tatarowiczj@battelle.org

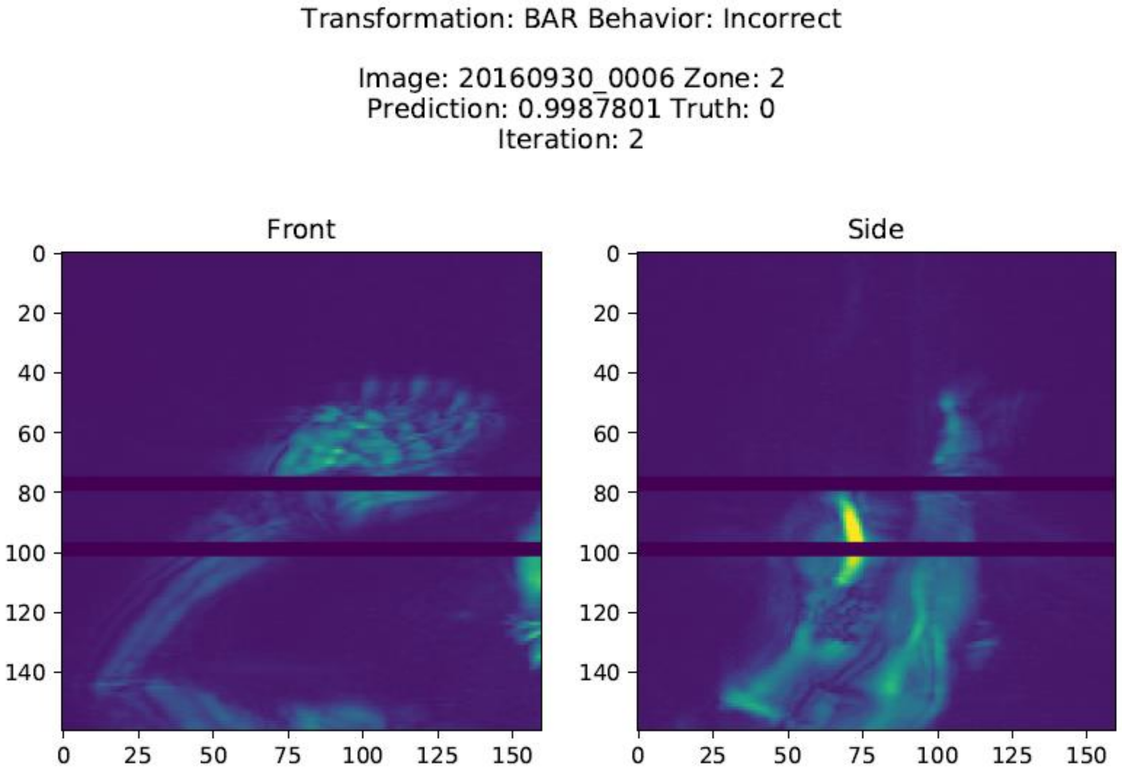
U.S. DEPARTMENT OF HOMELAND SECURITY

Homeland Security

Science and Technology

Image Augmentations: Dead Detector

False Positive

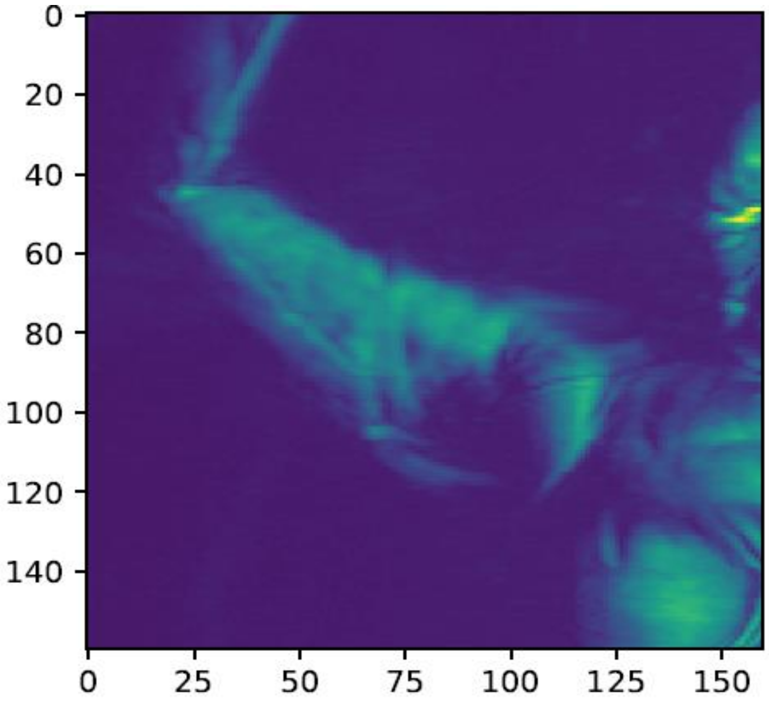


Transformation: BAR Behavior: Incorrect

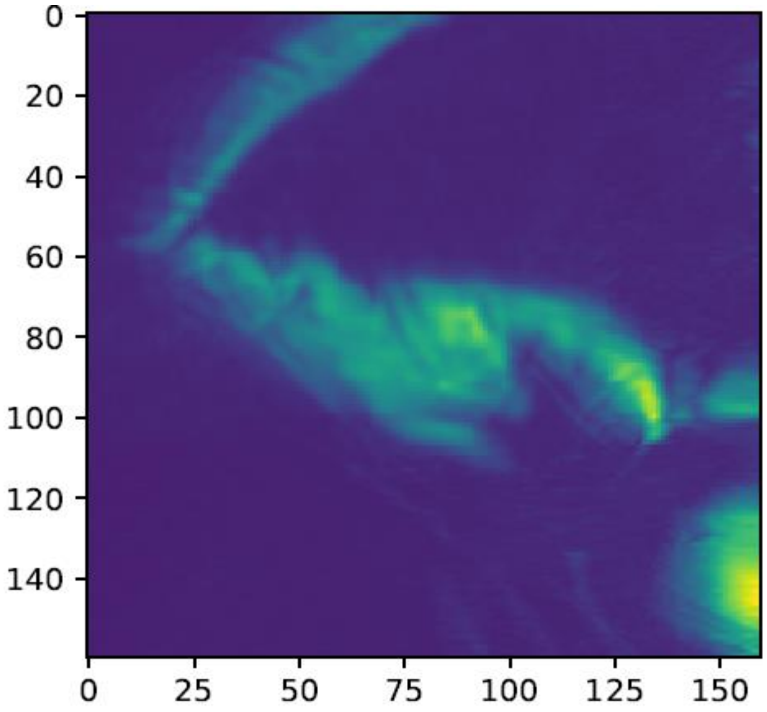
Image: 20160930_0006 Zone: 2
Prediction: 0.9987801 Truth: 0
Iteration: 2

Real vs. Synthetic Blur Comparison

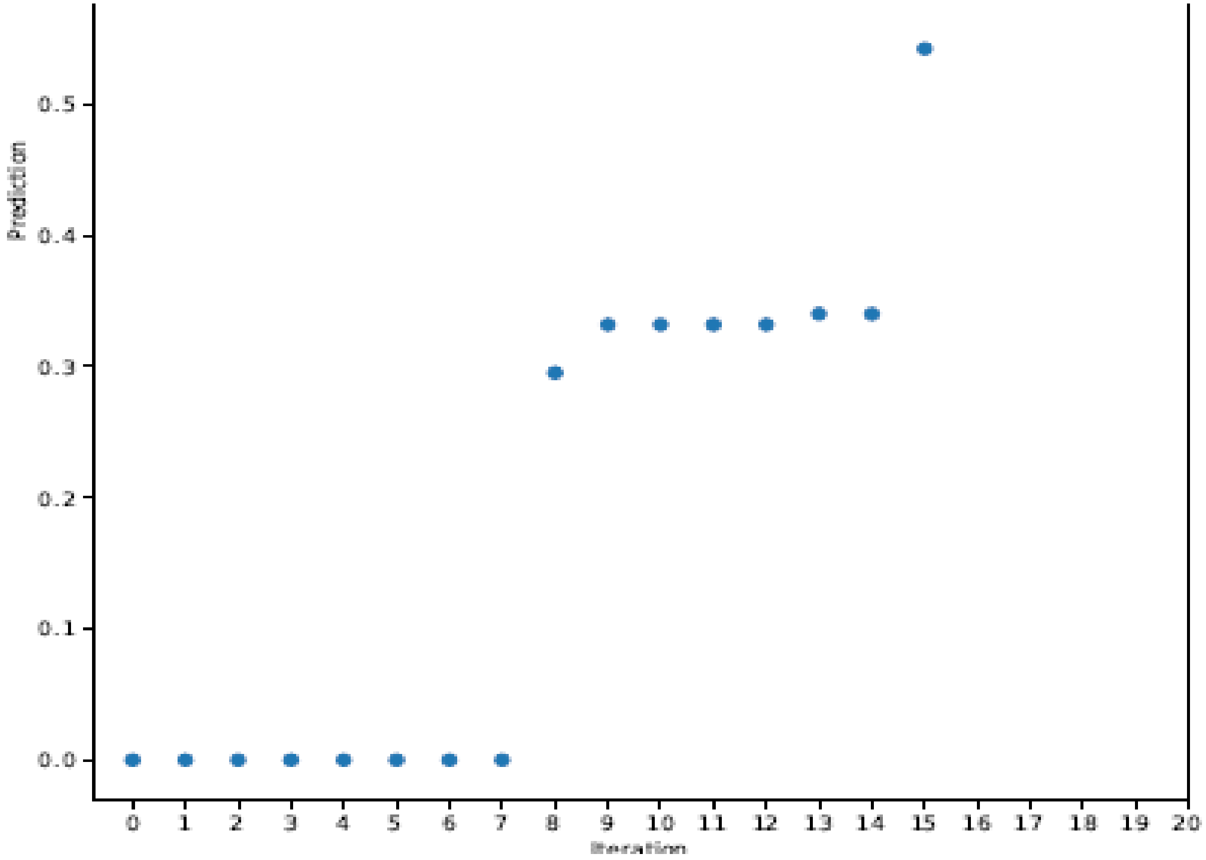
Real Blur



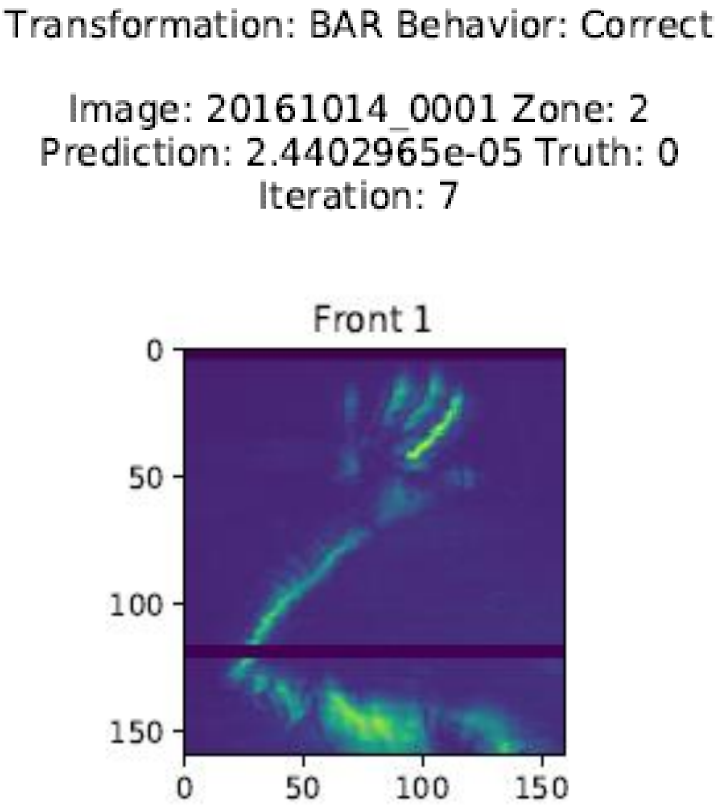
Synthetic Blur



Data Collection: Scatter Plots

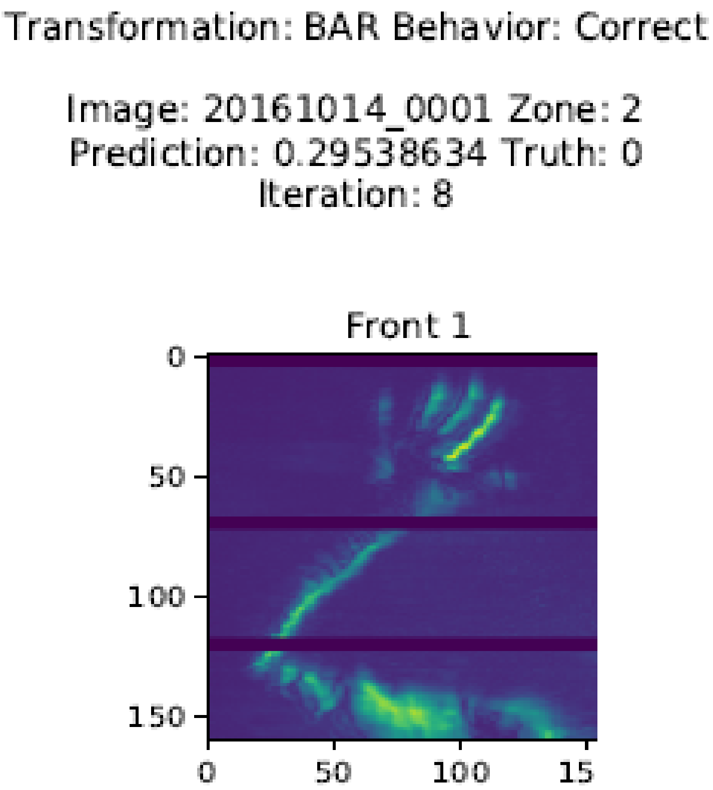


Significant Jumps in Scatter



Transformation: BAR Behavior: Correct

Image: 20161014_0001 Zone: 2
Prediction: 2.4402965e-05 Truth: 0
Iteration: 7



Transformation: BAR Behavior: Correct

Image: 20161014_0001 Zone: 2
Prediction: 0.29538634 Truth: 0
Iteration: 8