Interpretable AI for Deep-Learning-Based Meteorological Applications

Eric B. Wendoloski, Ingrid C. Guch

The Aerospace Corporation

Abstract

The number of machine-learning (ML) applications has surged within the meteorological community over the last several years. This surge includes the development and application of numerous ML techniques to improve forecasting as well as physical models while reducing computational cost. Given the vast trove of available satellite-based weather imagery and the gridded structure of many meteorological datasets, deep-learning (DL) methods for providing predictions and diagnostics for numerous subdomains are experiencing increased adoption. However, full adoption will require forecasters and decision makers to interpret why model output is produced given the input, especially if the output has implications for human well-being. Interpreting DL models can be especially difficult due to their complex architectures, and such models are often treated as black boxes. This work examines contemporary methods for assessing the interpretability of a convolutional neural network (CNN) trained to predict tropical cyclone (TC) intensity based on available satellite imagery, primarily in the IR band. CNNs excel at distilling images into the most important feature abstractions for developing functional associations between input and model output. The goal of this work is to assess whether such a DL architecture is capable of learning physically relevant abstractions for the problem at hand. We will describe and apply interpretability methods to the TC intensity CNN model to assess the importance of physical concepts to final predictions. We will also assess the traceability of predictions across the learned network. Additionally, methods for assessing model vulnerability to adversarial inputs are explored.

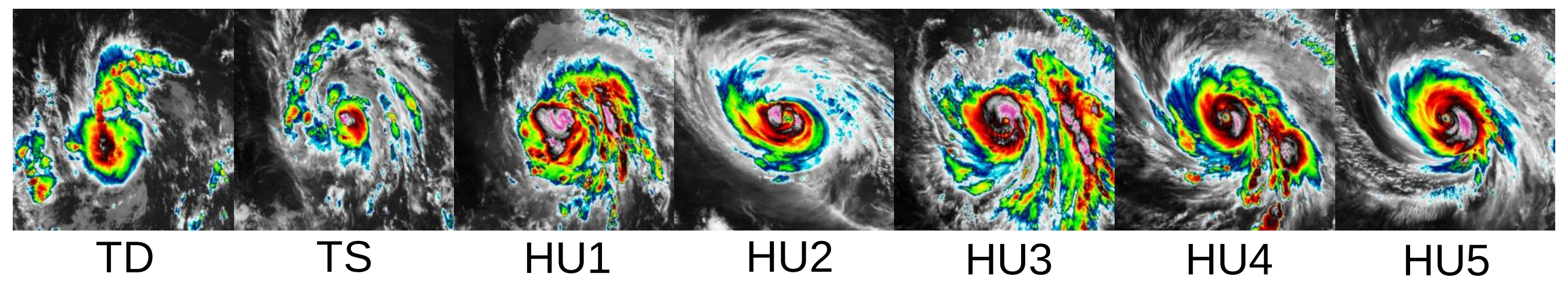
Introduction

- Difficulty explaining origins of artificial-intelligence (AI) prediction a key barrier to adoption
- Machine-learning (ML) models (basis for AI) often perceived as black boxes
  - Notion prevalent in deep-learning
  - O(10M-100M+) trainable parameters
  - Filters within layers trained to distill relevant features that map to some output
- Operational users require explainable decision processes / knowledge of model vulnerabilities
  - Predictions informing life and death decisions must be traceable
  - Predictions of physical processes should depend on physically relevant features
  - Vulnerabilities may be exploited to alter predictions
- Explainable AI techniques capable of conveying DL model behavior / vulnerabilities
  - Demonstrate physical features/concepts the model relates to its predictions
  - Provide understanding of model vulnerability to adversarial inputs

Demonstration Convolutional Neural Net (CNN)

- Model intended to highlight application of methodologies
- Trained to categorize tropical cyclones (TC; TD – Cat. 5)
  - Null class of randomly pulled regions included
- TC Data: 2017-18 Atlantic / Eastern Pacific TCs
- CNN Inputs: GOES-E 11.2 um band
  - Image chips centered on TC center of circulation
  - ~14000 images from 71 TCs

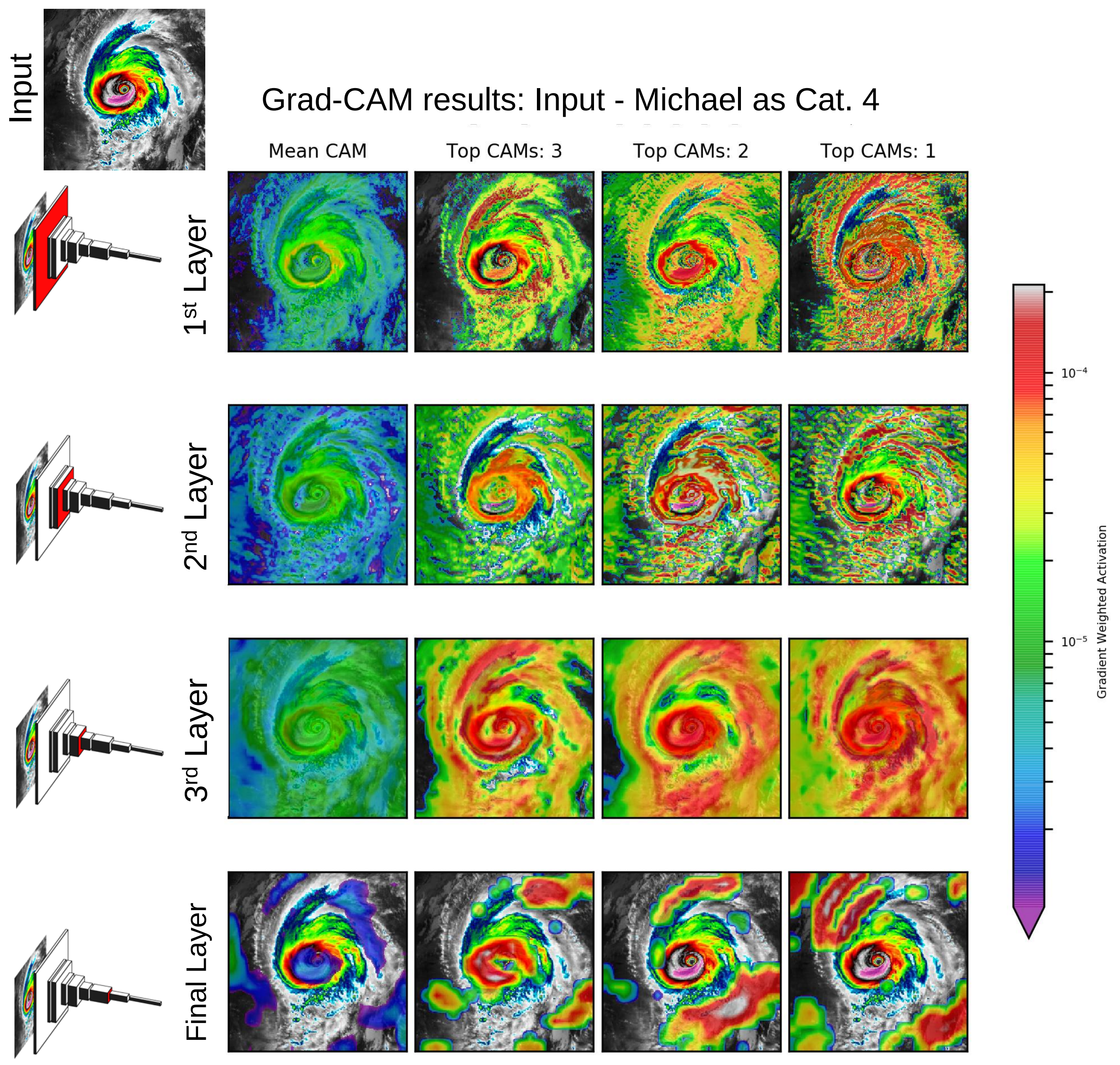
Maria TD-HU5 image chip examples



Explainable AI Methodologies

Gradient Weighted Class Activation Mapping (Grad-CAM)

- Visual explanation for CNN decisions
  - Identifies input image pixels most important to class prediction
- Method (Selvaraju et al. 2017)
  - Run image through CNN & gather layer activations
  - Compute gradient of predicted score for class of interest w.r.t. activations
  - Average the gradients – one avg. gradient for each filter in the layer
  - Weight activations by respective gradient
  - Results aggregated as layer mean or viewed by filter
  - Results local to input image

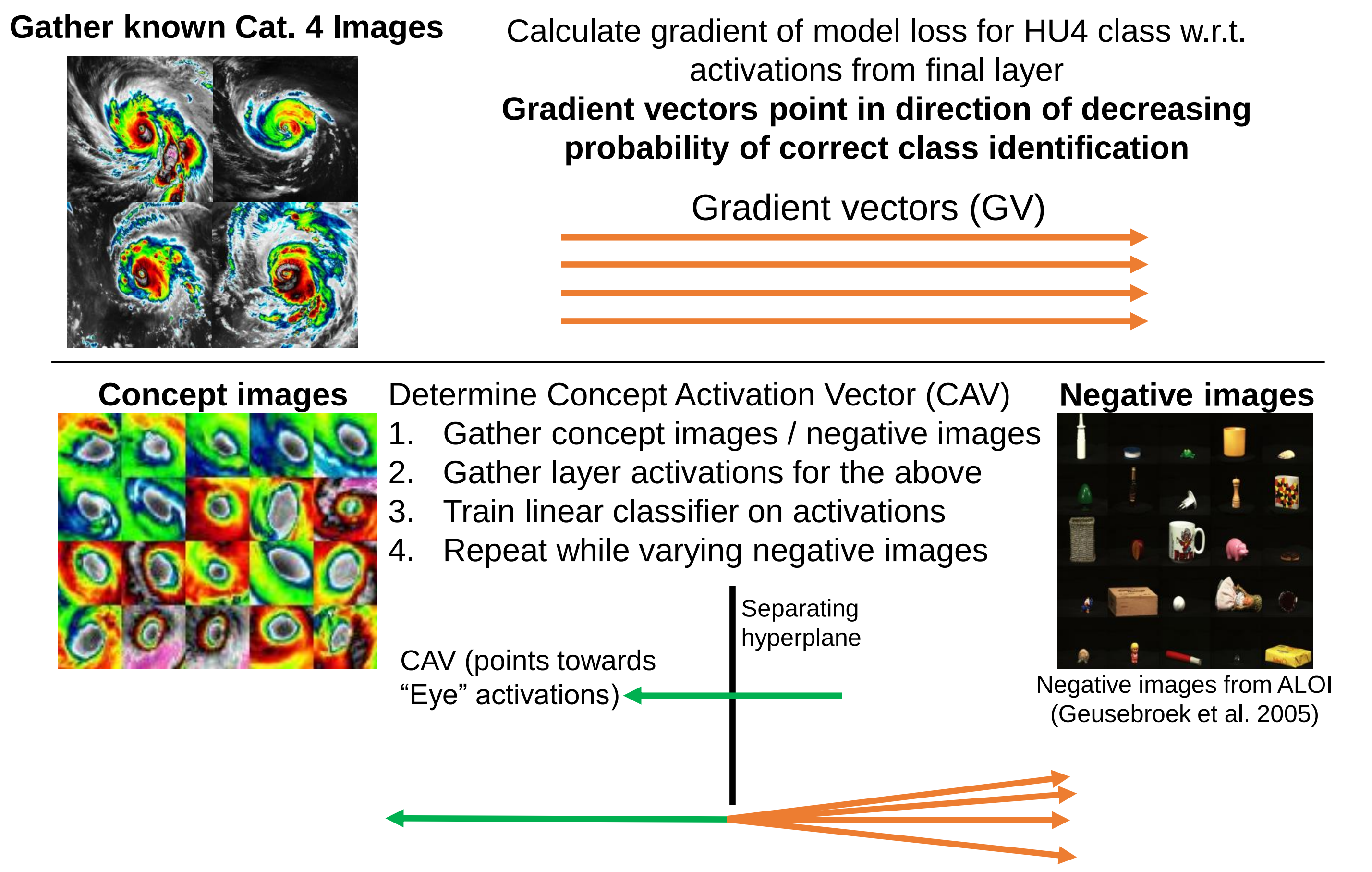

Grad-CAM results: Input - Michael as Cat. 4

Explainable AI Methodologies Cont.

Testing with Concept Activation Vectors (TCAV)

- Determine significance of user-defined concept for predictions of a given class (Kim et al. 2018)

Importance of Eye Structure to Cat. 4 Prediction



Gather known Cat. 4 Images

Calculate gradient of model loss for HU4 class w.r.t. activations from final layer

Gradient vectors point in direction of decreasing probability of correct class identification

Gradient vectors (GV)

Concept images

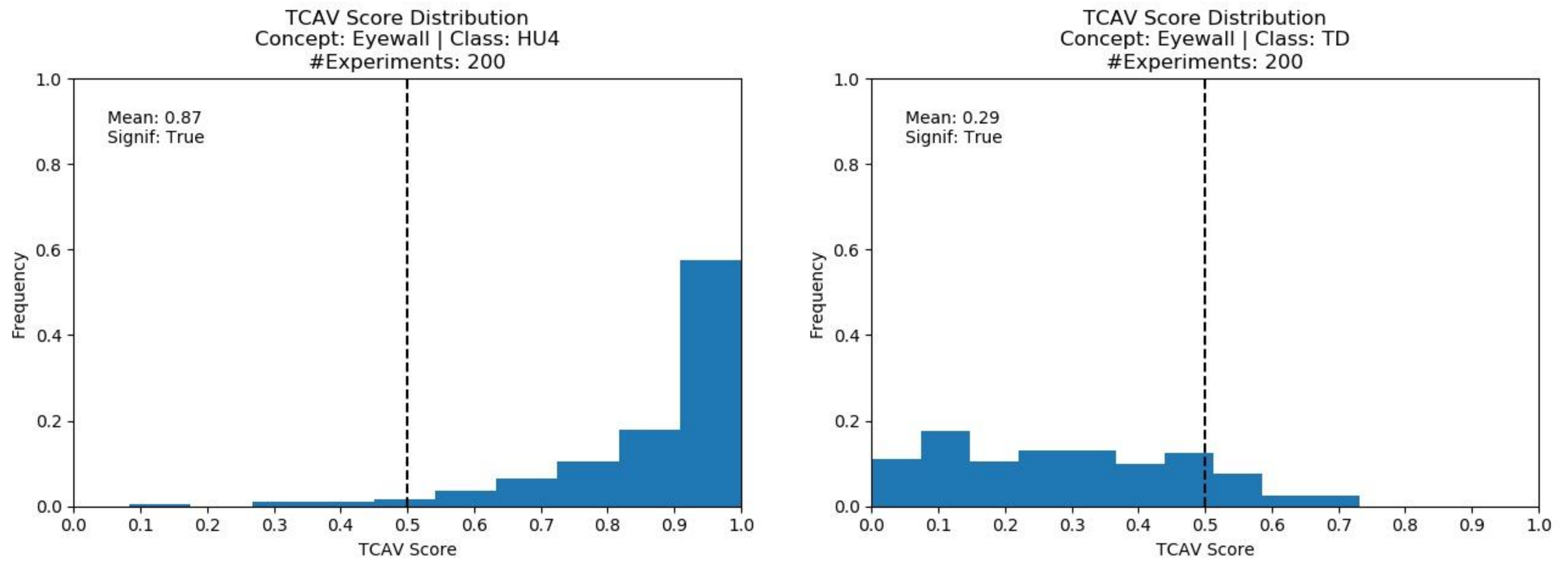
Determine Concept Activation Vector (CAV)

1. Gather concept images / negative images
2. Gather layer activations for the above
3. Train linear classifier on activations
4. Repeat while varying negative images

Negative images

Negative images from ALOI (Geusebroek et al. 2005)

- CAV tending to point in opposite direction of GVs tends to point in direction of increasing probability of correct class identification



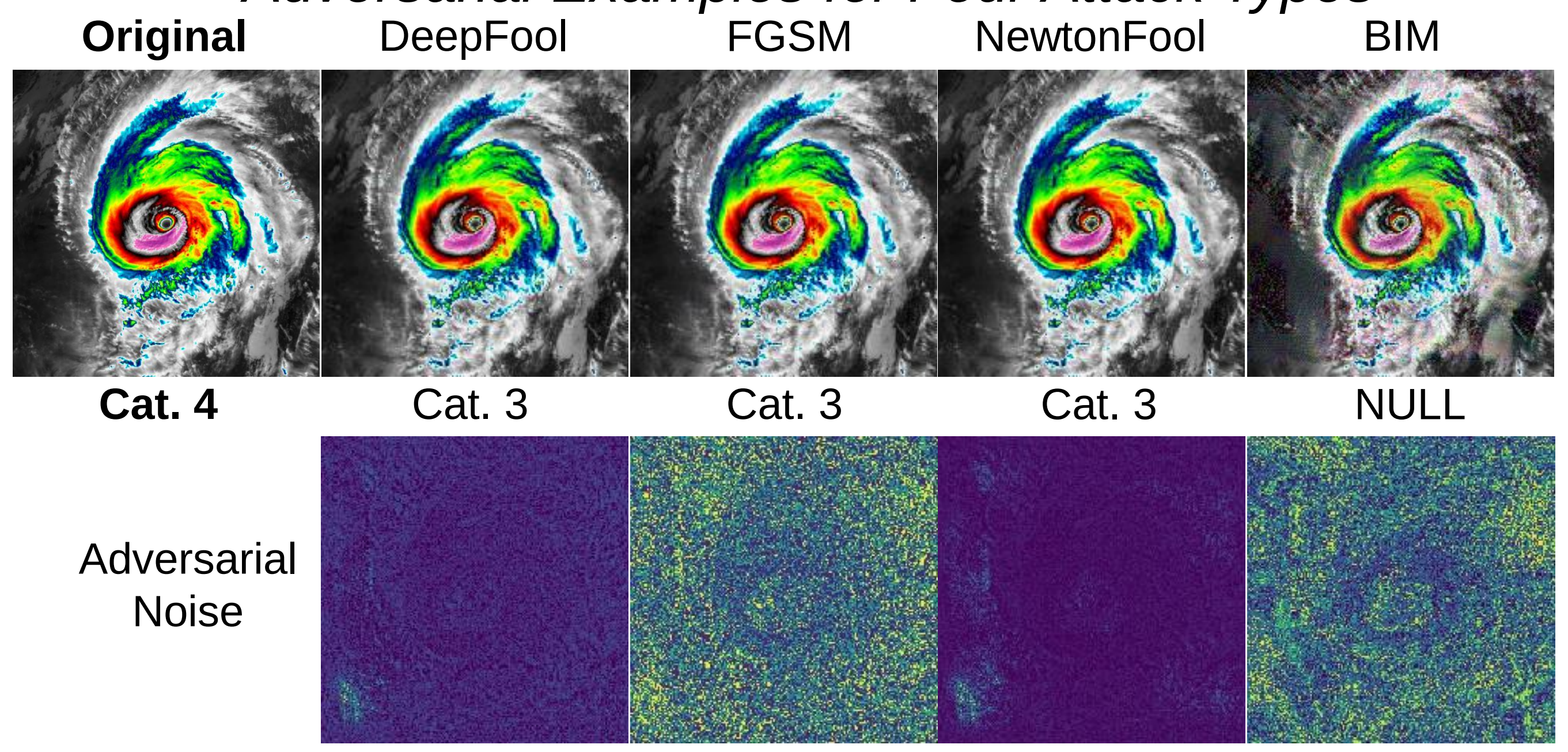
- Concept important if CAVs point opposite of >50% of GVs
- 87% of Cat. 4 GVs in opposite direction of “Eye” CAV
  - *Eye concept significantly important to prediction of Cat. 4 TCs*
- 29% of TD GVs in opposite direction of “Eye” CAV

Robustness to Adversarial Inputs

- Adversarial attacks may cause erroneous model inference
  - Occurs at testing or deployment stage
  - Targeted   – misguide to specific class
  - Untargeted – misguide to arbitrary class
- Adversarial noise added to images causes misclassification
  - Noise often imperceptible to humans
  - Noise exposes and exploits flaws in model decision function
  - Model-to-model transferability possible (Papernot et al. 2016)
  - Security risk when attacker has no access to victim model
- Operational use of CNNs requires knowledge of vulnerabilities
  - Ability to detect / screen / remediate adversarial inputs

Robustness to Adversarial Inputs Cont.

Adversarial Examples for Four Attack Types



Attack success rates on ~120 correctly predicted Michael images

Attack	Top-1 Accuracy (%)
Original (clean)	100
DeepFool (Moosavi-Dezfooli et al. 2016)	40.7
FGSM (Goodfellow et al 2015)	0.81
NewtonFool (Jang et al. 2017)	0.00
BIM (Kurakin et al. 2016)	38.2

Attack mitigation strategies

- Model hardening: train on adversarial imagery; shown to offer regularization
- Filter noise through data preprocessing at test/deployment

Summary

- Contemporary methods capable of highlighting CNN decision process / concepts important to prediction
- ML models vulnerable to adversarial input
  - Mitigation of attack risk possible via hardening & filtering
- Future efforts
  - Extension of explainable approaches to regression
  - Visualization of decision-boundary improvements
  - General methodology to increase attack-agnostic robustness

References

Geusebroek, J, G. Burghouts, and A. Smeulders, 2005: The Amsterdam library of object images, *International Journal of Computer Vision*, **61**, 103-112.

Goodfellow, I, J. Shlens, and C. Szegedy, 2015: Explaining and harnessing adversarial examples. https://arxiv.org/abs/1412.6572v3

Jang, U., X. Wu, and S. Jha, 2017: Objective metrics and gradient descent algorithms for adversarial examples in machine learning. https://doi.org/10.1145/3134600.3134635

Kim, B., M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, 2018:Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). https://arxiv.org/abs/1611.11279v5

Kurakin, A., I. Goodfellow, and S. Bengio, 2016: Adversarial machine learning at scale. https://arxiv.org/abs/1807.012396

Landsea, C. W. and J. L. Franklin, 2013: Atlantic Hurricane Database Uncertainty and Presentation of a New Database Format. Mon. Wea. Rev., 141, 3576-3592.

Moosavi-Dezfooli, S., A. Fawzi, and P. Frossard, 2016: DeepFool: a simple and accurate method to fool deep neural networks. https://arxiv.org/abs/1511.04599

Nicolae M.I., and Coauthors, 2018: Adversarial robustness toolbox v0.3.0. https://arxiv.org/abs/1807.01069

Papernot, N., P.D. McDaniel, and I.J. Goodfellow, 2016: Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv:1605.07277

RAMMB, 2018: Real-time tropical cyclone products – 2018 Season. Accessed 12 November 2018, http://rammb.cira.colostate.edu/products/tc_realtime/season.asp?storm_season=2018

Selvaraju, R.R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, 2017: Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. https://arxiv.org/abs/1610.02391v3

Approved For Public Release. OTR201900632.