Robust Neural Networks

Part 5: Conclusions and Future Directions

Key Takeaways

Role of Gradients

- **Robustness** under distributional shift in domains, environments, and adversaries are **challenges** for neural networks
  - **Gradients at Inference** provide a **holistic solution** to the above challenges
- **Gradients** can help **traverse** through a trained and unknown **manifold**
  - They approximate **Fisher Information** on the projection
  - They can be **manipulated** by providing **contrast** classes
  - They can be used to construct **localized contrastive** manifolds
  - They provide **implicit knowledge** about **all classes**, when only **one data** point is available at inference
- Gradients are useful in a number of **Image Understanding** applications
  - Highlighting features of the current prediction as well as **counterfactual** data and **contrastive** classes
  - Providing **directional information** in anomaly detection
  - **Quantifying uncertainty** for out-of-distribution, corruption, and adversarial detection
  - Providing **expectancy mismatch** for human vision related applications

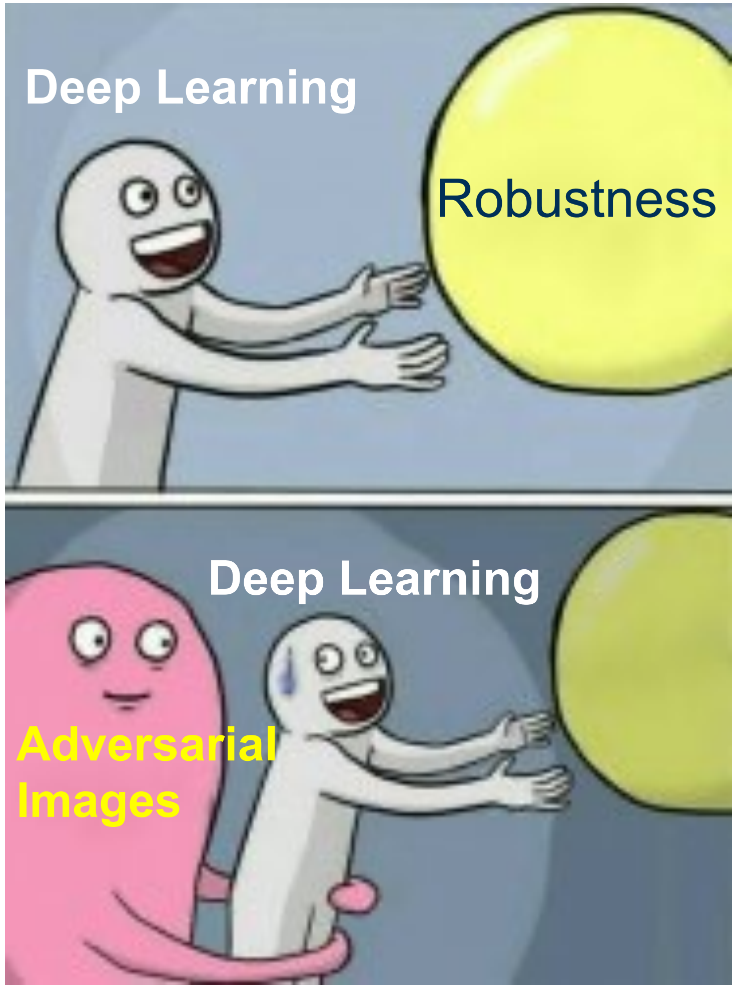
Future Directions

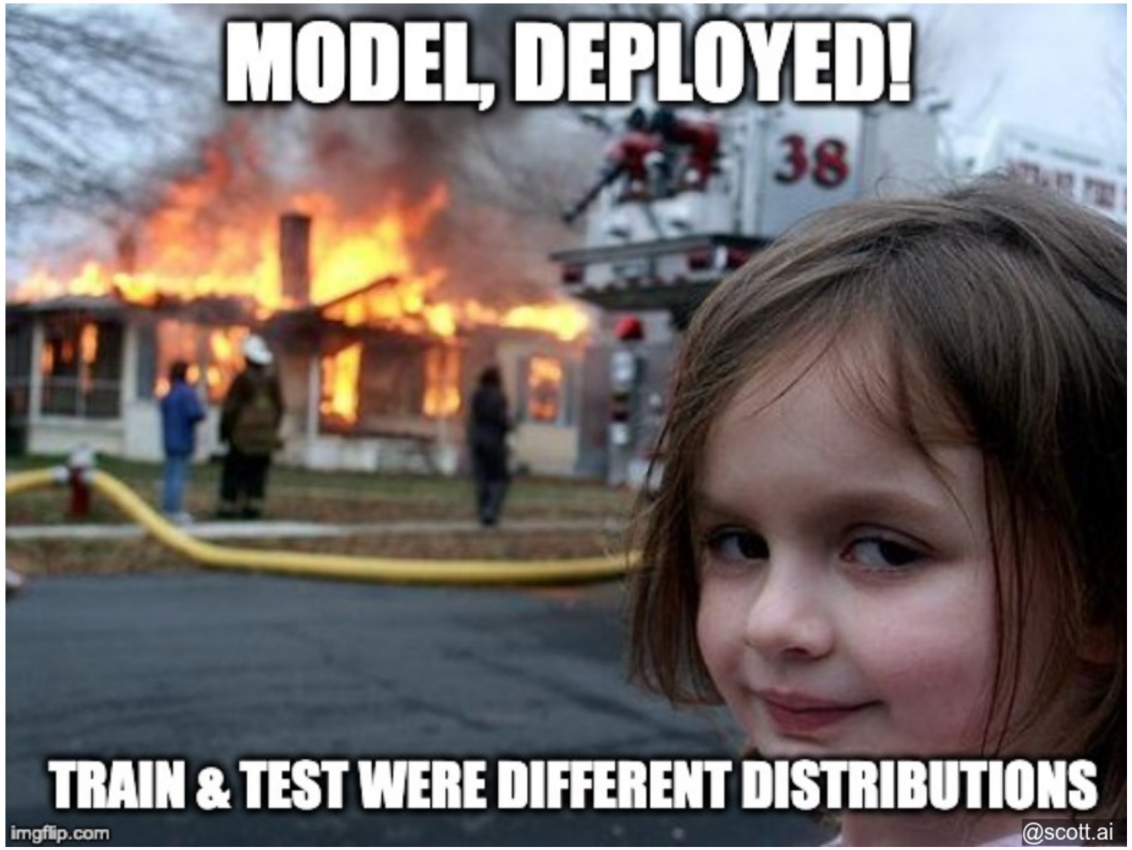
Research at Inference Stage

- **Test Time Augmentation (TTA) Research**
  - Multiple augmentations of data are passed through the network at inference
  - Research is in designing the best augmentations
- **Active Inference**
  - Utilize the knowledge in Neural Networks to ***ask it to ask us***
  - Neural networks ask for the best augmentation of the data point given that one data point at inference
- **Uncertainty in Explainability, Label Interpretation, and Trust quantification**
  - Uncertainty research has to expand beyond model and data uncertainty
  - In some applications within medical and seismic communities, there is no agreed upon label for data. Uncertainty in label interpretation is its own research
- **Test-time Interventions for AI alignment**
  - Human interventions at test time to alter the decision-making process is essential trustworthy AI
  - Further research in intelligently involving experts in a non end-to-end framework is required

Memes to Wrap it Up

Robustness at Inference





Cannot depend on training to construct robust models

References

Gradient representations for Robustness, OOD, Anomaly, Novelty, and Adversarial Detection

- **Gradients for robustness against noise:** M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022
- **Gradients for adversarial, OOD, corruption detection:** J. Lee, M. Prabhushankar, and G. AlRegib, "Gradient-Based Adversarial and Out-of-Distribution Detection," in *International Conference on Machine Learning (ICML) Workshop on New Frontiers in Adversarial Machine Learning*, Baltimore, MD, Jul. 2022.
- **Gradients for Open set recognition:** Lee, Jinsol, and Ghassan AlRegib. "Open-Set Recognition With Gradient-Based Representations." *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021.
- **GradCon for Anomaly Detection:** Kwon, G., Prabhushankar, M., Temel, D., & AlRegib, G. (2020, August). Backpropagated gradient representations for anomaly detection. In *European Conference on Computer Vision* (pp. 206-226). Springer, Cham.
- **Gradients for adversarial, OOD, corruption detection :** J. Lee, C. Lehman, M. Prabhushankar, and G. AlRegib, "Probing the Purview of Neural Networks via Gradient Analysis," in IEEE Access, Mar. 21 2023.
- **Gradients for Novelty Detection:** Kwon, G., Prabhushankar, M., Temel, D., & AlRegib, G. (2020, October). Novelty detection through model-based characterization of neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3179-3183). IEEE.
- **Gradient-based Image Quality Assessment:** G. Kwon*, M. Prabhushankar*, D. Temel, and G. AlRegib, "Distorted Representation Space Characterization Through Backpropagated Gradients," in *IEEE International Conference on Image Processing (ICIP)*, Taipei, Taiwan, Sep. 2019.

Explainability in Neural Networks

- **Explanatory paradigms:** AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks: Towards relevant and contextual explanations. *IEEE Signal Processing Magazine*, *39*(4), 59-72.
- **Contrastive Explanations:** Prabhushankar, M., Kwon, G., Temel, D., & AlRegib, G. (2020, October). Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3289-3293). IEEE.
- **Explainabilty in Limited Label Settings:** M. Prabhushankar, and G. AlRegib, "Extracting Causal Visual Features for Limited Label Classification," in IEEE International Conference on Image Processing (ICIP), Sept. 2021.
- **Explainabilty through Expectancy-Mismatch:** M. Prabhushankar and G. AlRegib, "Stochastic Surprisal: An Inferential Measurement of Free Energy in Neural Networks," in Frontiers in Neuroscience, Perception Science, Volume 17, Feb. 09 2023.

References

Self Supervised Learning

- **Weakly supervised Contrastive Learning:** K. Kokilepersaud, S. Trejo Corona, M. Prabhushankar, G. AlRegib, C. Wykoff, "Clinically Labeled Contrastive Learning for OCT Biomarker Classification," in IEEE Journal of Biomedical and Health Informatics, 2023, May. 15 2023.
- **Contrastive Learning for Fisheye Images**: K. Kokilepersaud, M. Prabhushankar, Y. Yarici, G. AlRegib, and A. Parchami, "Exploiting the Distortion-Semantic Interaction in Fisheye Data," in *Open Journal of Signals Processing*, Apr. 28 2023.
- **Contrastive Learning for Severity Detection:** K. Kokilepersaud, M. Prabhushankar, G. AlRegib, S. Trejo Corona, C. Wykoff, "Gradient Based Labeling for Biomarker Classification in OCT," in *IEEE International Conference on Image Processing (ICIP)*, Bordeaux, France, Oct. 16-19 2022
- **Contrastive Learning for Seismic Images:** K. Kokilepersaud, M. Prabhushankar, and G. AlRegib, "Volumetric Supervised Contrastive Learning for Seismic Semantic Segmentation," in *International Meeting for Applied Geoscience & Energy (IMAGE)*, Houston, TX, , Aug. 28-Sept. 1 2022

Human Vision and Behavior Prediction

- **Pedestrian Trajectory Prediction:** C. Zhou, G. AlRegib, A. Parchami, and K. Singh, "TrajPRed: Trajectory Prediction With Region-Based Relation Learning," *IEEE Transactions on Intelligent Transportation Systems*, submitted on Dec. 28 2022.
- **Human Visual Saliency in trained Neural Nets:** Y. Sun, M. Prabhushankar, and G. AlRegib, "Implicit Saliency in Deep Neural Networks," in *IEEE International Conference on Image Processing (ICIP)*, Abu Dhabi, United Arab Emirates, Oct. 2020.
- **Human Image Quality Assessment:** D. Temel, M. Prabhushankar and G. AlRegib, ”UNIQUE: Unsupervised Image Quality Estimation,” in IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1414-1418, Oct. 2016.

Open-source Datasets to assess Robustness

- **CURE-TSD:** D. Temel, M-H. Chen, and G. AlRegib, "Traffic Sign Detection Under Challenging Conditions: A Deeper Look Into Performance Variations and Spectral Characteristics," in *IEEE Transactions on Intelligent Transportation Systems*, Jul. 2019
- **CURE-TSR:** D. Temel, G. Kwon*, M. Prabhushankar*, and G. AlRegib, "CURE-TSR: Challenging Unreal and Real Environments for Traffic Sign Recognition," in *Advances in Neural Information Processing Systems (NIPS) Workshop on Machine Learning for Intelligent Transportation Systems*, Long Beach, CA, Dec. 2017
- **CURE-OR:** D. Temel*, J. Lee*, and G. AlRegib, "CURE-OR: Challenging Unreal and Real Environments for Object Recognition," in *IEEE International Conference on Machine Learning and Applications (ICMLA)*, Orlando, FL, Dec. 2018

References

Active Learning

- **Active Learning and Training with High Information Content:** R. Benkert, M. Prabhushankar, G. AlRegib, A. Parchami, and E. Corona, "Gaussian Switch Sampling: A Second Order Approach to Active Learning," in IEEE Transactions on Artificial Intelligence (TAI), Feb. 05 2023
- **Active Learning Dataset on vision and LIDAR data:** Y. Logan, R. Benkert, C. Zhou, K. Kokilepersaud, M. Prabhushankar, G. AlRegib, K. Singh, E. Corona and A. Parchami, "FOCAL: A Cost-Aware Video Dataset for Active Learning," IEEE Transactions on Circuits and Systems for Video Technology, submitted on Apr. 29 2023
- **Active Learning on OOD data:** R. Benkert, M. Prabhushankar, and G. AlRegib, "Forgetful Active Learning With Switch Events: Efficient Sampling for Out-of-Distribution Data," in *IEEE International Conference on Image Processing (ICIP)*, Bordeaux, France, Oct. 16-19 2022
- **Active Learning for Biomedical Images**: Y. Logan, R. Benkert, A. Mustafa, G. Kwon, G. AlRegib, "Patient Aware Active Learning for Fine-Grained OCT Classification," in *IEEE International Conference on Image Processing (ICIP)*, Bordeaux, France, Oct. 16-19 2022

Uncertainty Estimation

- **Gradient-based Uncertainty:** J. Lee and G. AlRegib, "Gradients as a Measure of Uncertainty in Neural Networks," in *IEEE International Conference on Image Processing (ICIP)*, Abu Dhabi, United Arab Emirates, Oct. 2020
- **Gradient-based Visual Uncertainty:** M. Prabhushankar, and G. AlRegib, "VOICE: Variance of Induced Contrastive Explanations to Quantify Uncertainty in Neural Network Interpretability," *Journal of Selected Topics in Signal Processing*, submitted on Aug. 27, 2023.
- **Uncertainty Visualization in Seismic Images:** R. Benkert, M. Prabhushankar, and G. AlRegib, "Reliable Uncertainty Estimation for Seismic Interpretation With Prediction Switches," in *International Meeting for Applied Geoscience & Energy (IMAGE)*, Houston, TX, , Aug. 28-Sept. 1 2022.
- **Uncertainty and Disagreements in Label Annotations:** C. Zhou, M. Prabhushankar, and G. AlRegib, "On the Ramifications of Human Label Uncertainty," in *NeurIPS 2022 Workshop on Human in the Loop Learning*, Oct. 27 2022
- **Uncertainty in Saliency Estimation:** T. Alshawi, Z. Long, and G. AlRegib, "Unsupervised Uncertainty Estimation Using Spatiotemporal Cues in Video Saliency Detection," in *IEEE Transactions on Image Processing*, vol. 27, pp. 2818-2827, Jun. 2018.

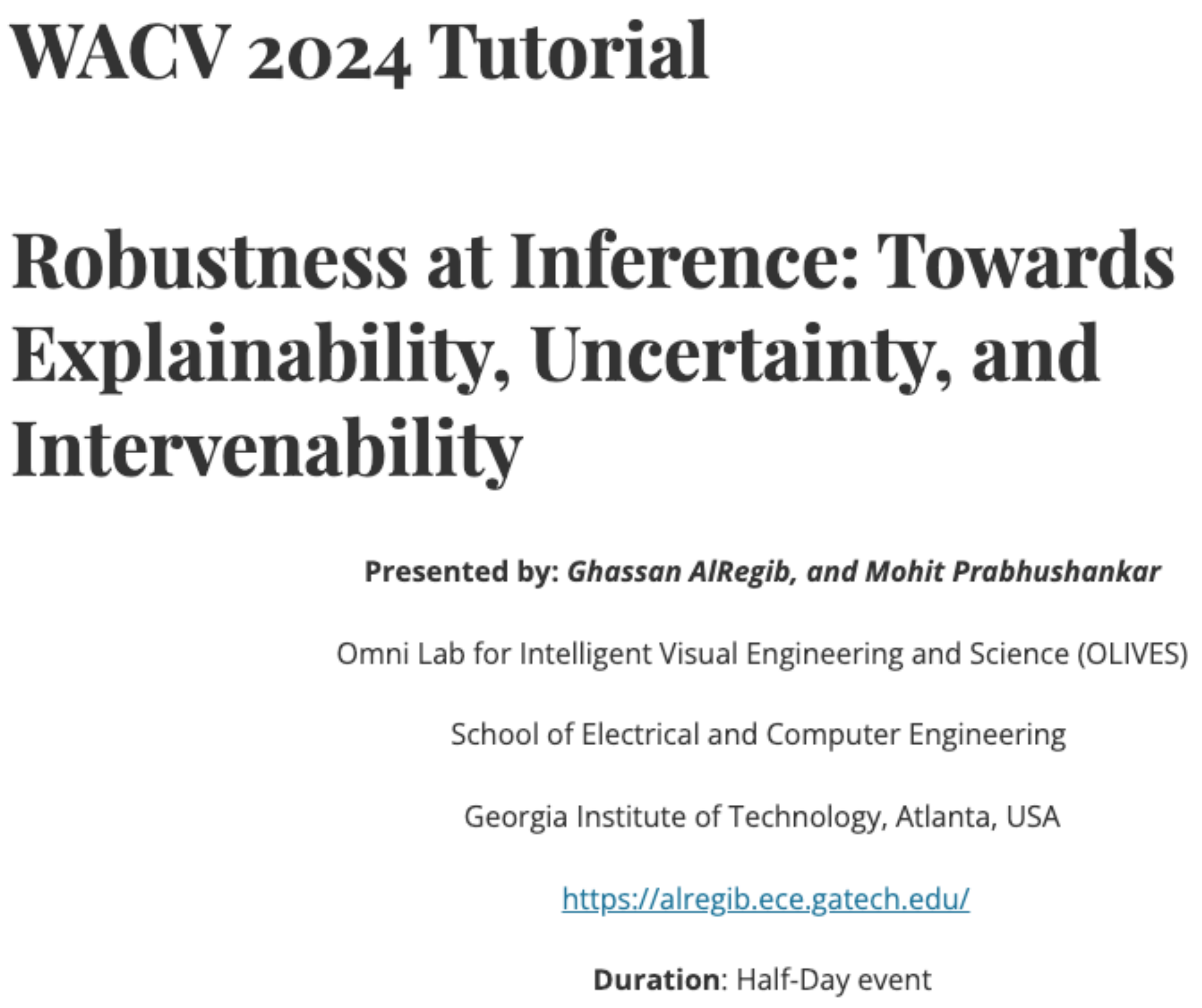
Tutorial Materials

Accessible Online

SCAN ME

https://alregib.ece.gatech.edu/wacv-2024-tutorial/

{alregib, mohit.p}@gatech.edu



WACV 2024 Tutorial

Robustness at Inference: Towards Explainability, Uncertainty, and Intervenability

Presented by: ***Ghassan AlRegib, and Mohit Prabhushankar***

Omni Lab for Intelligent Visual Engineering and Science (OLIVES)

School of Electrical and Computer Engineering

Georgia Institute of Technology, Atlanta, USA

https://alregib.ece.gatech.edu/

Duration: Half-Day event