Utilizing Gaze Information from Labelers: Can Bridging the Gap Between What Humans and Machines Fixate On Improve Model Performance and Explainability?

Air Force Research Laboratory | Information Directorate | Core Technical Competencies

Autonomy, Command and Control, and Decision Support (AC2)
Connectivity and Dissemination (CAD)
Cyber Science and Technology (CYB)
Processing and Exploitation (PEX)

INFORMATION INSTITUTE MISSION: *Strengthen and expand information technology research, develop collaborative relationships, and increase research emphasis in areas of information technologies for the Information Directorate.*

BACKGROUND:

- **Limited Data:** Very limited data may require expert knowledge to analyze and annotate.
- **Model Explainability**: Explainability and trust are crucial for users making decisions.
- **Manual Annotations**: Hand-labeled annotations of salient areas in an input image may not reflect the brain's full decision-making process.

METHODS:

- Dataset from [Ehinger 2009] contains ~1000 images with target-present (human) and target-absent (no humans) labels
- Dataset includes gaze fixation information for 14 labelers for each image
- [Boyd et al. 2022] suggests blurring non-salient regions of input images may increase performance
- [Boyd et al. 2023] suggests implementing a CAM loss to encourage the machine to focus on salient regions

RESULTS:

- An an increase in classification accuracy was observed after the implementation of CAM loss derived from dataset labeler gaze.
- An improvement to model explainability was also identified through a decrease in CAM loss, suggesting the neural network is more closely aligned with the human labelers in regards to the regions it focuses on.

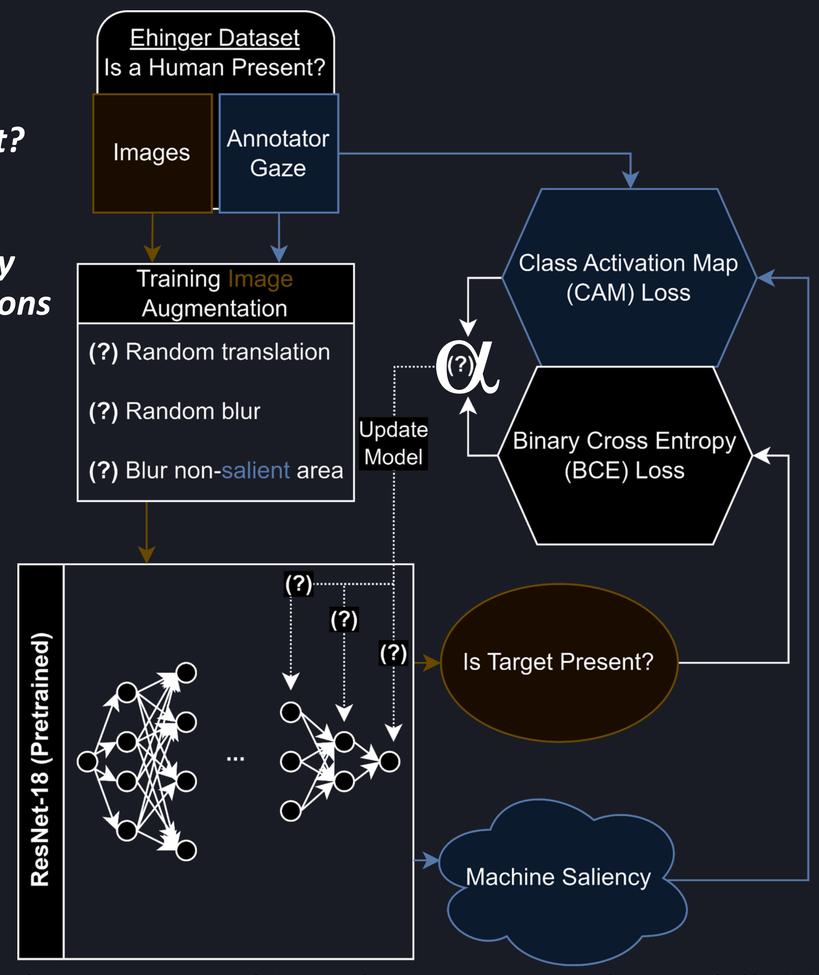
Kevin Barkevich
Ph.D Student
Rochester Institute of Technology
PerForM Lab

Supervised at AFRL by Dr. Erin Tripp

Hand-labeled saliency maps produced by dataset labelers have been shown to improve the quality of neural networks.

Can saliency information extracted from eye-tracking data provide the same benefit?

Can the resulting networks reliably demonstrate that they are focusing on the same regions as the human labelers?



Dataset: [Ehinger et al. 2009]
- "Human" search task
- Strongly guided search
  - [Mack & Palmeri 2011]
- Fixations from 14 labelers

(?) Will blurring the areas of a training image people gazed more infrequently at improve the model's performance?

(?) Will balancing the loss from human-to-machine classification with the loss from human-to-machine saliency improve the model's performance and explainability? [A/Alpha Parameter]

(?) How much of a pre-trained model should be unfrozen for the best performance?

This research was made possible through funds made available through the Air Force Office of Scientific Research and in part by the Air Force Research Laboratory, Information Directorate, Contract Number FA8750-20-3-1003.

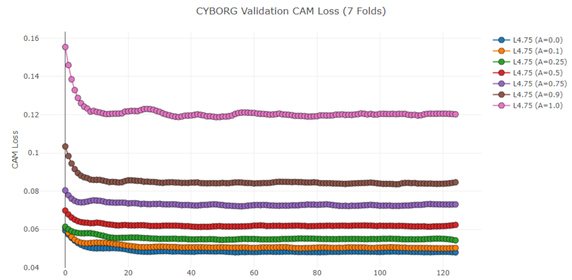
Approved for Public Release, Distribution Unlimited, AFRL-2024-5681

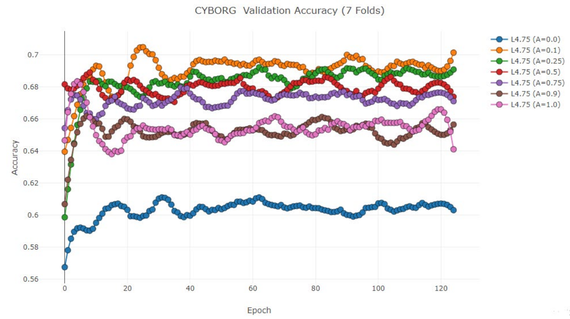
Possible Training Augmentations:
- Random cropping/translation
- **Fixation-based blurring** (gaussian blurring regions labelers did not fixate on)

Possible Alpha values (balance between the machine fixation--based CAM loss and the classification task-based BCE loss):
- 1.0, 0.9, 0.75, 0.5, 0.25, 0.1, 0.0

Possible pre-trained layers to unfreeze:
- Layer 4 (all or part)
- Layers 4 & 3 & 2

Top Left: A target-present input image.
Top Right: A target-present input image with non-fixated areas blurred.
Bottom Left: A target-absent input image with non-fixated areas blurred.





REFERENCES:
Aidan Boyd, Kevin W. Bowyer, and Adam Czajka. 2022. Human-Aided Saliency Maps Improve Generalization of Deep Learning. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2735–2744.

Aidan Boyd, Patrick Tinsley, Kevin W. Bowyer, and Adam Czajka. 2023. CYBORG: Blending Human Saliency Into the Loss Improves Deep Learning-Based Synthetic Face Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 6108–6117.

Ehinger KA, Hidalgo-Sotelo B, Torralba A, Oliva A. Modeling Search for People in 900 Scenes: A combined source model of eye guidance. Vis cogn. 2009 Aug 1;17(6-7):945-978. doi: 10.1080/13506280902834720. PMID: 20011676; PMCID: PMC2790194.