HARMONIZING HUMAN ATTENTION AND DEEP NEURAL NETWORKS:

Gaze-Weighted Attention for Object Recognition

Andrew Shin

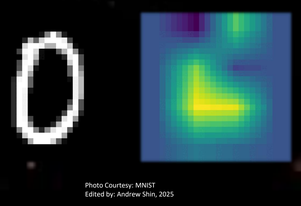
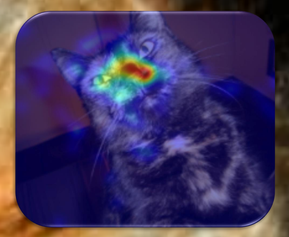


Photo Courtesy: MNIST
Edited by: Andrew Shin, 2025



INTRODUCTION - PROBLEM

Deep neural networks (DNNs) have achieved remarkable accuracy in image classification but often exhibit attention mechanisms that stray far from human visual processing. Furthermore, DNNs have grown increasingly via raw computational power, but not through an increase in intellectual abilities. This misalignment leads to brittle decision-making and reduced interpretability. ***This research investigates the potential of a neural harmonizer integrating human gaze data into DNN training to improve interpretability and accuracy.***

Deep Neural Networks (DNNs) are powerful but lack human-like attention

- AI models often fixate on irrelevant features in images.
- This leads to misinterpretation and reduced trust in AI.



Current AI struggles with explainability and alignment with human vision.

- Traditional models prioritize high-frequency patterns rather than meaningful objects.
- Example: In medical imaging, AI may focus on artifacts (labels, corners) instead of clinically relevant regions.



Photo Courtesy: Nature; F.Pasa Edited by: Andrew Shin, 2025

HYPOTHESIS

It is hypothesized that harmonizing DNN (Resnet50, ViT) attention with human visual saliency will create models that are more aligned with human cognition and achieve comparable or superior classification performance, increasing transparency of model decision-making.
This research will provide strong evidence that integrating human cognitive patterns into deep learning improves both explainability and task performance.

ABSTRACT

Standard DNNs do not align well with human vision. This research shows that the Neural Harmonizer breaks this misalignment, significantly improving normalized performance. This is strong evidence that aligning DNN feature importance with human vision improves both interpretability and performance in image classification.

To test this, a modified ResNet50 and Vision Transformer (ViT) focused on a harmonizer loss function is utilized, penalizing deviations between model-generated saliency maps and human gaze maps, utilizing Spearman rank correlation and Gaussian Pyramids. The models were evaluated on a dataset comprising ten object categories, each with corresponding human gaze heatmaps.

At the end, this research will show that the Neural Harmonizer trained on Eye Gaze Data shows better alignment, better metrics, increased interpretability, and a potential for broader applications.

OBJECTIVES

Research Goal

- Develop a **harmonization framework** that integrates **human gaze patterns** into **deep neural network (DNN) training**, aligning AI decision-making with human cognition.
- Improve **classification accuracy** while enhancing **interpretability**, ensuring models **prioritize features humans find meaningful** rather than relying on spurious correlations.
- Validate this approach across **multiple object categories**, demonstrating applicability **to real-world AI systems.**
- Verify the hypothesis of emplacing of our biological evolution to DNNs in increasing their **intellectual capability** in a transparent manner

Human Gaze Fixations:
Localize eye gaze locations into saliency representations from .mat files.

Neural Harmonizer:
λ value management for harmonization loss balance. *Spearman rank correlation + Gaussian Pyramids*

Improved Interpretability:
Validate regarding DNN explanation techniques, metrics, convolutional layers…

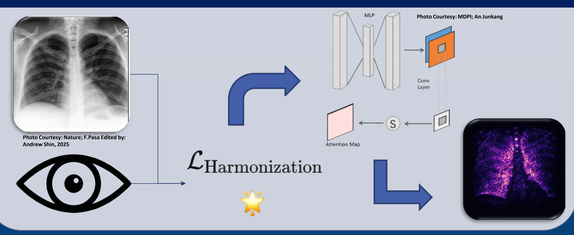
Claims

1. Primary Endpoint: Harmonized models achieve higher accuracy and interpretability, reducing errors in image classification.

2. Secondary Endpoint: The harmonization loss function ensures models focus on relevant regions, mimicking human perception. They will consist of different learning filters from original models.

3. Data Validation: Compare saliency maps before and after harmonization, observe convolutional layers beyond those shown through saliency maps, metrics, etc.

VISUAL FRAMING



METHODOLOGY

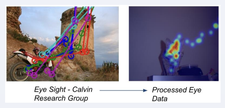
MATERIALS

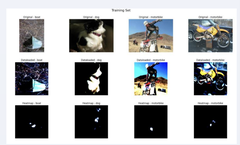
Software & Hardware:

- **Python**
- **Kaggle**
- **TensorFlow & PyTorch**
- **NumPy & Pandas**
- **Matplotlib & Seaborn**
- **OpenCV**
- **GPU Servers (NVIDIA Tesla T4 x2 & P100)**

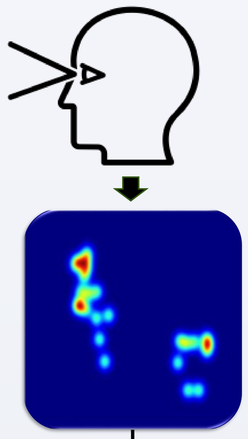
Datasets and Human Gaze Data:

- **Pascal Objects Eye Tracking (POET) v1.1:** Contained 10 object classes with corresponding human gaze heatmaps

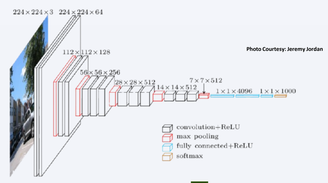




EYE GAZE MAP



REGULAR DNN



HARMONIZATION

$$\mathcal{L}_{\text{Harmonization}} = \lambda_1 \sum_{i}^{N} \left\| (z \circ \mathcal{P}_i \circ g(f_\theta, x))^+ - (z \circ \mathcal{P}_i(\phi_{\text{gaze}}))^+ \right\|_2 + \mathcal{L}_{CCE}(f_\theta, x, y) + \lambda_2 \sum_{i} \theta_i^2$$

Loss Function Modified from Serre Lab

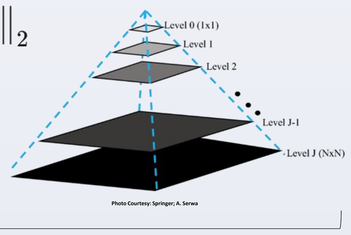


Photo Courtesy: Springer; A. Serwa

Human Gaze-Driven Attention Learning:
Extract human gaze heatmaps and standardize saliency distributions. Integrate eye-tracking data to enhance feature alignment in ***convolutional layers***

Creating the Neural Harmonizer:
1. ***Gaussian Pyramid:*** *features across scales for detail capturing (broad gazes + strict object boundaries)*
2. ***Harmonization Loss:*** *adjust* ***λ*** *parameters*

Model Performance & Evaluation:
Apply ***Spearman's Rank Correlation & IoU*** *to measure alignment between model-generated saliency maps and human gaze heatmaps.*

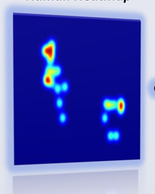
Key Notations

- $\mathcal{P}_i$ – Gaussian pyramid
- $g(f_\theta, x)$ – Saliency maps
- ϕ_{gaze} – Human gaze heatmaps
- z – Standardization
- $.^+$ – High-importance regions
- λ_1 – Harmonization loss weight
- λ_2 – Regularization weight

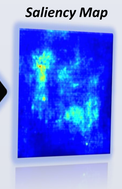
Object: Dog



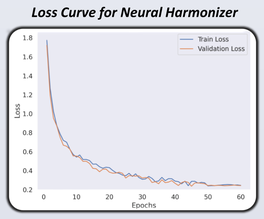
Human Heatmap



Saliency Map

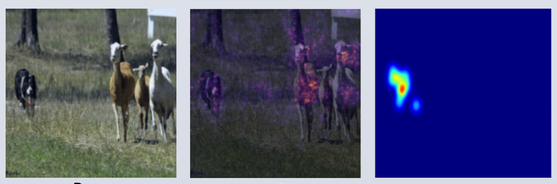


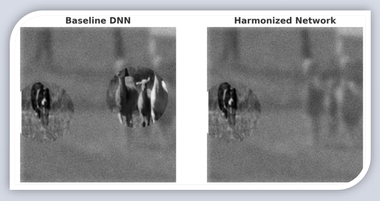
Loss Curve for Neural Harmonizer



RESULTS

Figure A. Human Feature Importance and Model Alignment





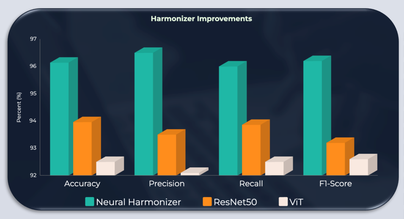
★ *Baseline DNN's focus on both objects during detection/classification, leading to potential errors. The harmonized network is inclined to choose based on human gaze.*

Activation Align - Interpretability

$$IoU_{k,c} = \frac{\sum |M_k(x) \cap L_c(x)|}{\sum |M_k(x) \cup L_c(x)|}$$

Numerator: Pixels that are in both the activation and human-labeled mask.
Denominator: Pixels in either mask

Figure B. Quantitative Performance Metrics



The bar chart comparing baseline and the ***Neural harmonizer*** *reveals a consistent improvement across all key performance metrics.*

Performance Summary
Harmonized models consistently outperform baseline models across all metrics.

Key Comparison
2.5% Accuracy Increase
93.8% vs. 96.1%

Real-World Implications
Potential Applications: Improved performance in medical imaging and critical tasks

Spearman Rank Correlation
Non-Linear & Monotonic

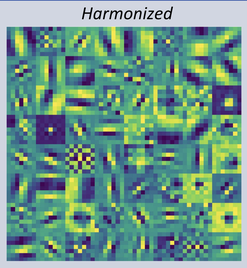
$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Overview:
- Spearman's rank correlation ranges from -1 to 1, where 1 indicates a perfect positive monotonic relationship

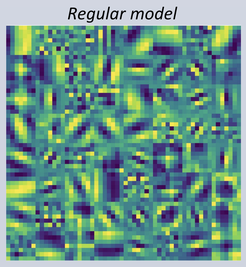
Harmonization Application:
- Used to quantify the alignment between human gaze fixation maps and DNN saliency maps.

Figure C. ResNet50 First Convolutional Layer

Harmonized



Regular model



★ *Cosine Similarity between Baseline and Harmonized Conv1 Filters: 0.07*

Harmonization significantly alters the model's perception of features, indicating a fundamental shift in learned representations at the lowest level.

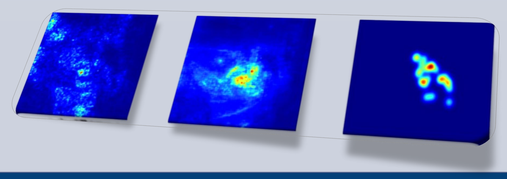
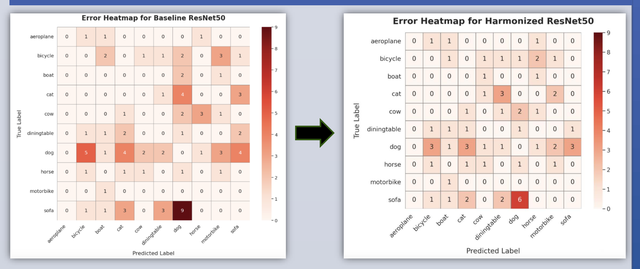


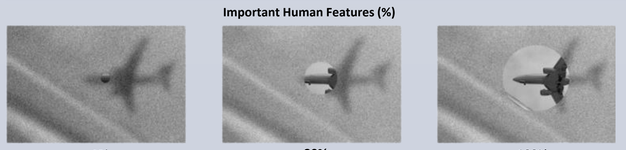
Figure D. Misclassification Analysis: Error Heatmaps



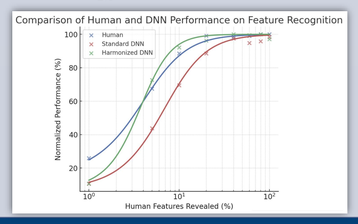
Initial - error counts reaching as high as 9 misclassified instances in a single category
Harmonizer - reduced misclassification in the "dog" vs. "cat" category from 8 errors to 5 errors

Figure E. Comparative Study on Human, Standard DNN, and Harmonized DNN Performance

Important Human Features (%)



1% **20%** **100%**



OBSERVATIONS:

- The experiment progressively increased the model's access to human features (1%, 20%, 100%)
- The harmonized model exhibits **progressive improvement in recognition performance as more human-relevant features are revealed**.

STATISTICAL ANALYSIS

Across both the final convolutional layer (from saliency maps) and first convolutional layer from visualization, the Neural Harmonizer is seen with clear deviations from baseline models; this leads to the increase in accuracy, precision, interpretability, and ultimately, its applicability to real-world usage. **Accuracy, for instance, has increased by 2.5% in the ViT model following harmonization.**

Spearman Correlation

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

$-1 < \rho < 1$

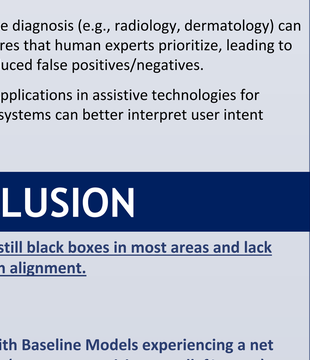
What it Measures:
- **"Is the model prioritizing those regions in the same way humans do?"**
- It captures how similar the importance *order of regions* is rather than pixel-wise overlap.

Before: ~30% Alignment → **After:** ~80% Alignment

Activation-Concept Alignment

$$IoU_{k,c} = \frac{\sum |M_k(x) \cap L_c(x)|}{\sum |M_k(x) \cup L_c(x)|}$$

What it Measures:
- **"Is the model looking in the same places as humans?"**
- IoU quantifies the direct pixel-wise overlap between human-labeled regions (Lc) and model activation maps (Mk).
- *Mainly for individual image analysis*



Violin Plot:
The Neural Harmonizer represents alignments of up to ~96% while the baseline model peaks at ~42% using Spearman correlation across the dataset

Layer Analysis: first & back

- The architectural layers of DNNs like ResNet50 offer some insight into where the model "sees", and if harmonization is making an impact.
- The first layers of the Neural Harmonizer and the regular model had a cosine similarity of **~0.07 (close to 0)**, indicating that the models look at different features even at its lowest layer
- The last layer representations are mainly shown through saliency maps like SmoothGrad, RISE Explanation, and more, and can clearly be overlapped to show that the **IoU of the Neural Harmonizer is significantly higher** than the baseline model
- These changes allow humans to better understand the working processes of AI and allow AI to be something aligned with how humans view the world, instead of being a complete **black box**.

SIGNIFICANCE

The findings of this research hold substantial implications across multiple domains, particularly in AI interpretability, human-centered AI, and real-world applications in medical imaging, autonomous systems, and human-computer interaction.

1. Advancing Explainable AI (XAI):

Current deep neural networks (DNNs) often function as "black boxes," making critical decisions without transparency. By integrating human gaze data into the training process, this study enhances the interpretability of AI models. The Neural Harmonizer ensures that AI decision-making aligns with human cognitive processes, fostering trust and transparency in AI-driven applications.

2. Bridging the Gap Between Human and Machine Vision

Traditional AI models rely on statistical correlations in data rather than human-like perception. This research demonstrates that aligning AI attention mechanisms with human visual focus significantly improves model interpretability without sacrificing accuracy. This is one of the first studies to showcase a scalable framework that effectively harmonizes AI and human vision in object recognition tasks.

3. Real-World Applications

- Medical Imaging: AI models for disease diagnosis (e.g., radiology, dermatology) can be trained to focus on the same features that human experts prioritize, leading to improved diagnostic accuracy and reduced false positives/negatives.
- Human-Computer Interaction (HCI): Applications in assistive technologies for individuals with disabilities, where AI systems can better interpret user intent based on gaze patterns.

CONCLUSION

Successfully Neural Networks are still black boxes in most areas and lack human alignment.

Harmonized models exhibited:

1. Improved classification accuracy, with Baseline Models experiencing a net increase in around ~2.5% in all metrics (accuracy, precision, recall, f1-score)

2. Greater alignment with human attention, as demonstrated by saliency maps, which showed harmonized models focusing on semantically relevant object regions.

3. Reduction in classification errors, particularly in visually ambiguous cases, as confirmed by error heatmaps.

4. Enhanced performance in feature recognition tasks, where harmonized models demonstrated higher alignment with human cognitive patterns compared to standard DNNs.

These findings suggest that human-inspired attention mechanisms can help mitigate the black-box nature of deep learning models by making their decision-making more transparent and interpretable. The improvement in accuracy further indicates that aligning AI with human perception does not come at the cost of performance but rather enhances it. This supports the broader premise that biologically inspired AI approaches can create models that are both effective and explainable.

FUTURE WORK

While this study demonstrated the effectiveness of harmonizing deep neural networks (DNNs) with human gaze, there are several areas for further improvement and expansion. One key direction for future work is scaling the dataset to include more diverse object categories and scene complexities. **The current dataset is limited to ten object classes**, and expanding it to broader, real-world datasets such as **MS COCO or ImageNet** could further validate the robustness of the harmonized models.

One critical area for improvement is **adaptive loss weighting**, which allows it to automatically adjust gaze alignment influence through confidence scores and uncertainty estimates. Furthermore, this study primarily focuses on image classification, but the harmonization framework could be extended to other vision tasks, such as **object detection, segmentation, and video analysis.**

There are several unresolved questions about **ethical matters and broad applicability** in gaze-guided artificial intelligence. Although human interpretation of attention provides important understanding it has limitations because it can **show bias and varies between cultures** and depends on specific tasks. When gaze harmonization functions in crucial environments such as healthcare or hiring practices special protection steps must be established to counteract discriminatory biases that can occur when gaze patterns do not function similarly in different population groups. **Future studies must investigate bias correction methods using adversarial learning and fairness-based gaze representation models to preserve transparent and equitable operation of harmonized AI systems among various user communities.**