Shared Interest...Sometimes: Understanding the Alignment between Human Perception, Vision Architectures, and Saliency Map Techniques

Katelyn Morrison, Ankita Mehra, Adam Perer
{kcmorris, aniktame}@andrew.cmu.edu, adamperer@cmu.edu

Abstract

Empirical studies have shown that attention-based architectures outperform traditional convolutional neural networks (CNN) in terms of accuracy and robustness. As a result, attention-based architectures are increasingly used in high-stakes domains such as radiology and wildlife conservation to aid in decision-making. However, understanding how attention-based architectures compare to CNNs regarding alignment with human perception is still under-explored. Previous studies exploring how vision architectures align with human perception evaluate a single architecture with multiple explainability techniques or multiple architectures with a single explainability technique. Through an empirical analysis, we investigate how two attention-based architectures and two CNNs for two saliency map techniques align with the ground truth for human perception on 100 images from an interpretability benchmark dataset. Using the Shared Interest metrics, we found that CNNs align more with human perception when using the XRAI saliency map technique. However, we found the opposite for Grad-CAM. We discuss the implications of our analysis for human-centered explainable AI and introduce directions for future work.

1. Introduction

Image classification techniques have rapidly advanced in the last few years since the introduction of attention-based architectures. With the popularity of attention-based architectures and their ability to outperform traditional CNNs, several domains, such as medical imaging and wildlife conservation, are starting to question whether they should start using attention-based architectures instead of CNNs [4, 6, 14, 15]. However, using these better-performing architectures in high-stakes domains does not eliminate the need or want for explanations.

Recent studies have qualitatively shown that saliency maps can be confusing and misleading (e.g., [21]). From our informal conversations with domain experts, it is evident that decision-makers in high-stakes domains are still relying on saliency maps to explain their models, as there are limited alternatives. For example, conservationists and biologists using a *WildBook* powered by WildMe.org's infrastructure can view saliency maps to understand individual species identification.

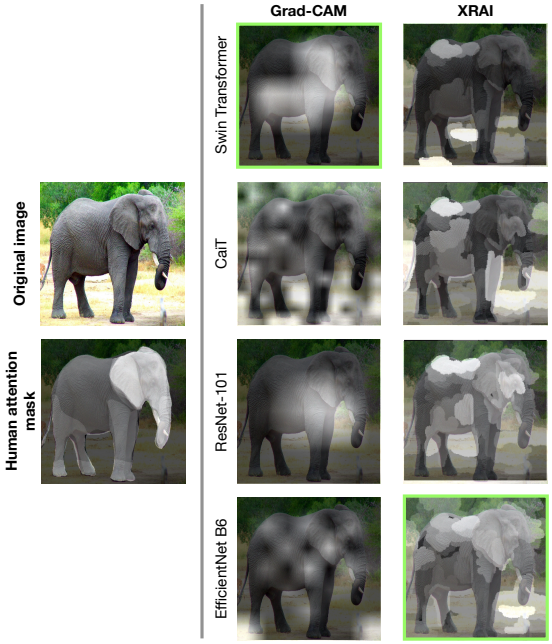

Figure 1. The human attention map and saliency maps for two techniques and four models are shown for the image of an African Elephant from ImageNet. The Swin Transformer had the highest IoU with the human attention map out of the Grad-CAM maps. The EfficientNet B6 had the highest IoU with the human attention map out of the XRAI maps. All models correctly predicted African Elephant.

Domain experts (i.e., conservationists) using an image classification model in their workflow are faced with many choices such as which architecture to choose and which saliency map technique to use. These choices are especially difficult given that explanations based on different techniques can disagree with one another [12, 20] or will highlight spurious patterns [18]. Previous work has tried

to simplify this choice by identifying techniques that align with human perception [2, 27]. However, these studies are limited in findings since they do not compare multiple architectures with multiple saliency map techniques.

We argue that it is necessary to do these empirical analyses with multiple architectures and multiple saliency techniques to understand [**RQ1**] which saliency map technique aligns more with human perception, [**RQ2**] which architecture aligns more with human perception, and [**RQ3**] to what extent the saliency technique and architecture choice play a role in how much "shared interest" there is with human perception. Through this empirical analysis, we address these research questions and contribute the following:

- We provide insight into how the saliency technique and architecture **choices are not trivial when prioritizing alignment with human perception**.
- We conduct the first empirical analysis that evaluates how **multiple image classification architectures** (i.e., attention-based architectures and CNNs) and **multiple saliency map techniques** align with human perception by using the Shared Interest metrics [3].

2. Related Work

Designing human-centered explainable AI (HCXAI) requires a deep understanding of human perception. We discuss several user studies that investigate how different image classification models or saliency map techniques align with human perception. We highlight how previous literature has contributed to HCXAI and distinguish our contributions to the field.

A recent study poses a similar question to ours: "*Are Convolutional Neural Networks or Transformers more like human vision?*" [24]. This empirical study investigates the "*observed error overlap*" of Google's Vision Transformer [8] to popular CNNs. They used Stylized-ImageNet [9] to evaluate the models' "*observed error overlap*". Ultimately, they found that the errors the Vision Transformer made were more aligned with human errors.

Instead of asking which architecture aligns more with human perception, one study asks if "*...Vision Transformers See Like Convolutional Neural Networks?*" [19] while another study asks if, "*transformers are more robust than CNNs*" [1]. Raghu et al. conducted rigorous quantitative analyses that lead them to discover that Vision Transformers (ViTs) are fundamentally different from CNNs.

While those two studies empirically evaluated transformers and CNNs, they use quantitative metrics that do not directly compare to human perception. Zhang et al. conducted a user study to evaluate how the saliency maps produced by SHAP for three different CNNs (Inception, ResNet, and VGG) align with human perception [27]. By collecting human input on regions of an image that are important to the class of the image, the authors compare the agreement between humans and the SHAP saliency map.

Similarly, Banerjee et al. seek to understand how human perception aligns with saliency maps produced by Grad-CAM for three different CNNs (VGG, EfficientNet, and ResNet) [2]. Unlike Zhang et al., Banerjee et al. use a saliency benchmark dataset [5].

Kapishnikov et al. propose a "region-based saliency method" for deep neural networks called XRAI [11]. XRAI is similar to Grad-CAM because it is also a gradient-based attribution method. Kapishnikov et al. quantitatively compare XRAI to Grad-CAM for an Inception and ResNet50 model, ultimately showing that it outperforms Grad-CAM [11]. We build off these contributions by comparing Grad-CAM to XRAI for attention-based models.

A recent study proposes a saliency technique that is designed specifically for attention-based architectures [17]. Using retinal disease images, they evaluate their approach algorithmically and qualitatively through a user study with four medical experts. From their user study, they observed that Grad-CAM was comparable to their technique for one dataset and significantly worse for another dataset.

3. Method

3.1. Pre-trained Models & Saliency Maps

Convolutional Neural Networks. We are using the EfficientNet B6 from the EfficientNet-PyTorch Python Library[1] and the ResNet-101 model from the `timm` library [25]. We chose these pre-trained models since they are more comparable to attention-based architectures in terms of the number of parameters and accuracy. ResNets and EfficientNets have also been used in previous empirical analyses [27].

Vision Transformers. The two attention-based architectures that we include in our empirical analysis are the small Swin Transformer with a patch size of 4 and window size of 7; and the small CaiT model [23]. We chose to look at these two models specifically because their smaller versions are more comparable to traditional convolutional neural networks in terms of the number of trainable parameters and accuracy. Both models are pre-trained and provided by the `timm` Python library [25].

Saliency Maps. While there are saliency map techniques that have been designed specifically for attention-based architectures, such as Focused Attention [17], we chose to use Grad-CAM [10, 22] and XRAI [11]. We chose Grad-CAM because it is a very popular technique for image classification and previous empirical studies include Grad-CAM in their analyses. We chose XRAI because it is a region-based saliency technique and is in the same family

[1] https://github.com/lukemelas/EfficientNet-PyTorch

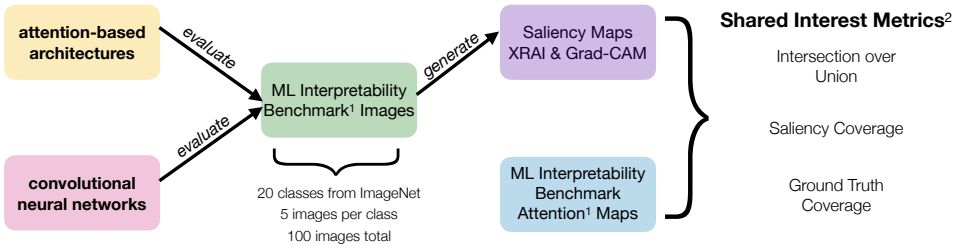

Figure 2. Overview of the design of our experiment. Using two attention-based architectures, two convolutional neural networks, and two saliency map techniques, we generated a total of eight saliency maps for 100 hundred images from ImageNet. Then, we use the Shared Interest metrics [3] to evaluate the saliency maps generated by the different techniques and architectures.

of attribution methods as Grad-CAM [11]. Grad-CAM was also a part of the main comparison for XRAI [11].

3.2. Dataset

We use the ML Interpretability Benchmark images for our empirical analysis [16]. There are a total of 20 classes from ImageNet and five images per class totaling 100 images. Along with the original images from ImageNet, the dataset includes aggregated human attention masks for each image. The attention masks are based on real-human data from a user study conducted by Mohseni et al. [16]. Throughout the paper, human attention masks and human perception are used interchangeably.

3.3. Experimental Design

As shown in Figure 2, we evaluate each of the four vision architectures on the 100 images from the ML Interpretability Benchmark dataset [16]. To generate the saliency maps for the CNNs, we use the last layer of each model. The last layer can be defined quite ambiguously and the Grad-CAM library [10] will average the CAM across the layers if it determines that multiple layers were passed in. For Resnet-101, we defined the last layer as `layer4[-1]`; for EfficientNet, `_blocks[-1]`; for Swin Transformer, `layers[-1].blocks[-1]`; and for CaiT, `blocks[-1]`.

Metrics. We leverage the three metrics proposed by Boggust et al., also known as the Shared Interest metrics [3]: intersection over union (IoU), ground truth coverage (GTC), and saliency coverage (SC). Specifically, we look at the IoU, GTC, and SC between the model's saliency map and human attention mask for each given image. Boggust et al. state that a high value for SC means that the model, "*relies almost exclusively on ground truth features*" and a high value for GTC indicates that, "*the ground truth features are the most relevant to the model's decision*".

4. Results

By evaluating and comparing different architectures and saliency map techniques to human attention masks using the Shared Interest metrics, we highlight when models, saliency maps, and human perception align. We present results for all of the Shared Interest metrics but only discuss the IoU results due to page limits.

4.1. RQ1: Shared Interest with Saliency Techniques

To understand if there is a saliency technique that generally aligns more with human perception than another, we calculate the mean IoU for Grad-CAM and the mean IoU for XRAI as if different models were not used. The distribution of these values for each saliency technique is shown at the top of Figure 3. The mean IoU for Grad-CAM is 0.192 and 0.145 for XRAI. A t-test between the two means shows that Grad-CAM is significantly different from XRAI ($p < 0.001$) and that we can conclude Grad-CAM aligns more with human perception than XRAI regardless of the architecture for the layer specifications that we used.

4.2. RQ2: Shared Interest with Architectures

If we don't consider the saliency map techniques, we can calculate the mean IoU for each model. As shown in the bottom boxplot in Figure 3, the CaiT model is the only model that has a significantly different IoU compared to the other three models ($p < 0.001$). However, the Swin-T, EfficientNet, and ResNet are not significantly different.

4.3. RQ3: Shared Interest with Technique and Architecture

Table 1 reports mean values for all images regardless of all four models predicting the same class. After filtering for when all four models predict the same exact class, the trends and significance remain the same.

A one-way ANOVA for the IoU for each model for Grad-CAM shows that the mean IoU between the four models is significantly different ($p < 0.001$). From a pairwise posthoc Tukey HSD test, we can conclude that the IoU for

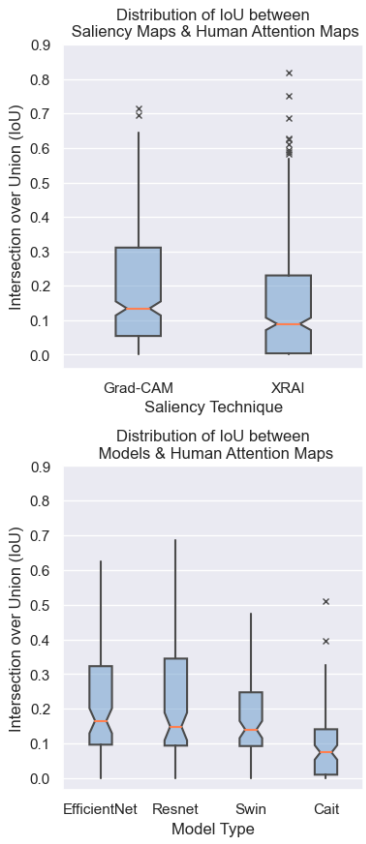

Figure 3. Two boxplots showing the distribution of IoU scores based on saliency technique (top) and model type (bottom).

all four models is significantly different from each other ($p < 0.01$) except the ResNet-101 and EfficientNet B6, and the ResNet-101 and Swin-T.

A one-way ANOVA for the IoU for each model for XRAI shows that the mean IoU between the four models is significantly different ($p < 0.001$). From a pairwise posthoc Tukey HSD test, we can conclude that there is a significant difference between the attention-based architectures and the CNNs ($p < 0.01$). We can also conclude that there is no significant difference between the EfficientNet B6 and ResNet-101, and the Swin-T and CaiT.

5. Discussion and Future Work

We found that when XRAI is paired with EfficientNet B6, it aligns more with human perception than Grad-CAM paired with EfficientNet B6. However, we observed the opposite trend for the ResNet-101. These trends do not align with those found by Kapishnikov et al. where XRAI outperformed Grad-CAM when paired with an Inception and ResNet50 model [11]. However, we observed Grad-CAM to be more aligned with human perception than XRAI when disregarding the model.

Comparison with Human Perception				
Explanation	Model	$\overline{IoU}$	$\overline{SC}$	$\overline{GTC}$
XRAI	Swin-T	0.079	0.255	0.109
	CaiT	0.082	0.271	0.103
	ResNet	0.197	0.435	0.253
	EfficientNet	**0.221**	**0.463**	**0.284**
Grad-CAM	Swin-T	**0.264**	**0.511**	**0.347**
	CaiT	0.101	0.273	0.142
	ResNet	0.216	0.466	0.287
	EfficientNet	0.189	0.411	0.264

Table 1. We report the mean for each shared interest metrics between the saliency map and the human attention masks.

We observed that saliency map techniques have shared interest with human perception... sometimes. It depends on the architecture being used, whether that be an attention-based architecture or a CNN. Human-centered explainable AI should take "shared interest...sometimes" into consideration when designing and evaluating techniques and architectures with human perception.

Based on our findings, it is evident that **choosing a saliency map technique is not trivial** when it comes to prioritizing interpretability. We encourage technical researchers to deeply understand what components of the saliency map techniques and architectures lead to increased alignment with human perception in order to design human-centered explainable AI.

We acknowledge this is preliminary work, making it difficult to generalize. For example, we only evaluated two saliency map techniques for four different models on 100 images. It is also difficult to capture the "ground truth" for human perception at scale. Therefore, to scale this empirical analysis, we plan to use a model to generate an approximation of human attention for a given image on a larger dataset such as the ImageNet Validation Set [7] or the Human Saliency Benchmark [26]. This can be achieved by using the Deep Gaze IIE [13] model by predicting where humans would look in an image.

6. Conclusion

We evaluated how attention-based architectures, CNNs, Grad-CAM, and XRAI align with human perception by using the Shared Interest metrics [3]. Using 100 images and attention maps from an interpretability benchmark [16], we are able to understand which architectures, saliency map techniques, and combination of the two align the most with human perception. We observed that while a saliency technique may align with human perception, this changes when paired with a particular type of architecture. As artificial intelligence continues growing, the design and evaluation of human-centered explainable AI need to adapt.

References

[1] Yutong Bai, Jieru Mei, Alan L Yuille, and Cihang Xie. Are transformers more robust than cnns? *Advances in Neural Information Processing Systems*, 34:26831–26843, 2021. 2

[2] Puja Banerjee, Siddhartha Raj, and Dr Rajesh P Barnwal. Machine versus human attention: A comparative study of different transfer learning models. In *Proceedings of the 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD)*, pages 136–136, 2023. 2

[3] Angie Boggust, Benjamin Hoover, Arvind Satyanarayan, and Hendrik Strobelt. Shared interest: Measuring human-ai alignment to identify recurring patterns in model behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2022. 2, 3, 4

[4] Yasamin Borhani, Javad Khoramdel, and Esmaeil Najafi. A deep learning based approach for automated plant disease classification using vision transformer. *Scientific Reports*, 12(1):11554, 2022. 1

[5] Ali Borji and Laurent Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581*, 2015. 2

[6] Marco Cantone, Claudio Marrocco, Francesco Tortorella, and Alessandro Bria. Convolutional networks and transformers for mammography classification: An experimental study. *Sensors*, 23(3):1229, 2023. 1

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2

[9] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 2

[10] Jacob Gildenblat and contributors. Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam, 2021. 2, 3

[11] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry. Xrai: Better attributions through regions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4948–4957, 2019. 2, 3, 4

[12] Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*, 2022. 1

[13] Akis Linardos, Matthias Kümmerer, Ori Press, and Matthias Bethge. Deepgaze iie: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12919–12928, 2021. 4

[14] Kangrui Lu, Yuanrun Xu, and Yige Yang. Comparison of the potential between transformer and cnn in image classification. In *ICMLCA 2021; 2nd International Conference on Machine Learning and Computer Application*, pages 1–6. VDE, 2021. 1

[15] Christos Matsoukas, Johan Fredin Haslum, Magnus Söderberg, and Kevin Smith. Is it time to replace cnns with transformers for medical images? *arXiv preprint arXiv:2108.09038*, 2021. 1

[16] Sina Mohseni, Jeremy E Block, and Eric D Ragan. Quantitative evaluation of machine learning explanations: A human-grounded benchmark. *arXiv preprint arXiv:1801.05075*, 2020. 3, 4

[17] Clément Playout, Renaud Duval, Marie Carole Boucher, and Farida Cheriet. Focused attention in transformers for interpretable classification of retinal images. *Medical Image Analysis*, 82:102608, 2022. 2

[18] Gregory Plumb, Marco Tulio Ribeiro, and Ameet Talwalkar. Finding and fixing spurious patterns with explanations. *arXiv preprint arXiv:2106.02112*, 2021. 1

[19] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021. 2

[20] Saumendu Roy, Gabriel Laberge, Banani Roy, Foutse Khomh, Amin Nikanjam, and Saikat Mondal. Why don't xai techniques agree? characterizing the disagreements between post-hoc explanations of defect predictions. In *2022 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 444–448. IEEE, 2022. 1

[21] Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, SQ Truong, CD Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G Blankenberg, AY Ng, et al. Deep learning saliency maps do not accurately highlight diagnostically relevant regions for medical image interpretation. *MedRxiv*, 2021. 1

[22] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2

[23] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 32–42, 2021. 2

[24] Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L Griffiths. Are convolutional neural networks or transformers more like human vision? *arXiv preprint arXiv:2105.07197*, 2021. 2

[25] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019. 2

[26] Yi Yang, Yueyuan Zheng, Didan Deng, Jindi Zhang, Yongxiang Huang, Yumeng Yang, Janet H Hsiao, and Caleb Chen Cao. Hsi: Human saliency imitator for benchmarking saliency-based model explanations. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 231–242, 2022. 4

[27] Zijian Zhang, Jaspreet Singh, Ujwal Gadiraju, and Avishek Anand. Dissonance between human and machine understanding. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019. 2