Towards Generating Human-Centered Saliency Maps without Significantly Sacrificing Accuracy

Vivek Aswal*, Gore Kao*, Seo Young Kim*, Katelyn Morrison*
Carnegie Mellon University, Pittsburgh, PA, USA
{vaswal, gorek, seoyoung, kcmorris}@andrew.cmu.edu

Abstract

As deep neural networks make significant advances in computer vision tasks, they are being deployed in several high-stakes domains. However, these models are not always semantically meaningful to humans as traditional interpretability techniques are quantitatively driven. Therefore, we explore how to generate saliency maps that are more similar to human attention without significantly sacrificing the model performance. We conduct an empirical study to understand how current object detection models compare to human centered saliency maps. Additionally, we present different data augmentation techniques such as Selective Erasing and Selective Inpainting along with the prevelant non-trivial transforms to evaluate the impact of human-centered data augmentation. With less than 3% mAP difference, we observe that data augmentations that are derived from predicted human attention improves the MAE and IoU between the model saliency and predicted attention. Visualization and more details are at here.

1. Introduction

Artificial intelligence (AI) is increasingly being built for and deployed in high-stakes domains such as prostate cancer detection from medical imagery [5], disaster relief efforts [7], and self-driving cars. However, these models are often treated as "black-boxes" and are not interpretable to people that collaborate or interact with them [1]. Therefore, the interpretability and accuracy of these models are equally important to calibrate decision-makers reliance on AI and improve human-AI collaboration.

With explainability and interpretability of AI becoming increasingly important, machine learning (ML) researchers designed a wide range of techniques from visualizing what the model has learned from an entire dataset, known as feature visualizations [15], to visualizing the pixels or regions of an image that activated a particular class prediction, known as class activation maps [18]. While these techniques are all derived quantitatively from the model, they are not always semantically meaningful to humans or even highlight the correct region in the image despite a correct prediction, known as spurious correlations [16]. As a result, human-computer interaction (HCI) researchers have been investigating how to make explainability techniques more interpretable by humans known as human-centered explainable AI (HCXAI) [10]. However, few works have explored HCXAI techniques for object detection models.

Our primary research questions are the following:

- How do current state-of-the-art object detection models compare to human attention?
- Can data augmentation techniques make saliency maps more similar to human attention without significantly sacrificing model accuracy?

We address these two research questions through two studies. First, we conduct a small empirical study to understand how current state-of-the-art object detection models compare to human attention. In the second study, we evaluate the impact of novel human-centered, data augmentations on deep neural networks (DNNs) saliency maps. Our novel contributions include (1) presenting two novel data augmentation techniques called *Selective Erasing* and *Selective Inpainting* that can be used for augmenting images for image classification and object detection models; and (2) evaluating the impact of different data augmentation techniques on saliency maps generated by Faster R-CNN.

2. Related Works

With novel interpretability techniques increasingly being developed, some researchers are taking a cognitive science approach to interpretability in order to understand how human attention compares to deep learning models. One study compares human attention to DNNs for segmentation, action recognition, and classification tasks [9].

Recently, Boyd et al., proposed a novel loss function that uses human annotations [4]. This loss function is de-

*All authors contributed equally and ordered alphabetically.

signed to penalize the model during training for generating saliency maps that are significantly different from the human saliency maps. The same authors just recently showed that human annotations can improve the generalization of a DNN [3]. For both of these studies, the authors had to collect ground truth annotations from human subjects in order to make the loss function which does not generalize well for other models or domains.

Instead of continuously having to collect human annotations, the MIT/Tuebingen Saliency Benchmark has designed a challenge for saliency prediction models [8]. For example, the DeepGazeIIE saliency prediction model is currently the best performing saliency prediction technique compared to gold standard metrics [12].

3. Methods

We conducted two studies in order to address our research questions. Below, we describe each study and how the second study uses results from the first study.

3.1. Empirical Study

We conducted an empirical study to gain an understanding of which state-of-the-art object detection models currently generate saliency maps similar to human attention. We evaluated and compared saliency maps generated by seven different object detection models available on PyTorch to human attention maps and predicted human eye-fixations. To obtain the human attention maps, we used the human attention maps for PASCAL2012 from the ML-Interpretability-Evaluation-Benchmark [13]. To obtain predicted human eye-fixations, we used the DeepGazeIIE saliency prediction model [11]. The specific state-of-the-art object detection models we evaluated and compared to the human attention maps and predicted eye-fixations include YOLOv5, Faster R-CNN with a ResNet-50 FPN backbone, SSD with a VGG backbone, SSD with a MobileNet backbone, Mask R-CNN, RetinaNet, and DETR.

3.1.1 Experiment Details

We generated a saliency map for every single image that had an associated human attention map (ground truth saliency) from the ML-Interpretability-Evaluation-Benchmark [13].

Each image was resized to 512 x 512 before being evaluated on by the model. The saliency map for the object detection model is generated using the EigenCAM method [14] from the PyTorch library for CAM methods [6]. The saliency map was generated using the last feature layer in each model (layer 4 in the backbone). Once the saliency map from the object detection model is generated, the mean absolute error (MAE) was calculated between the generated saliency map and the human attention map. The MAE is also calculated between the generated saliency map and the predicted human eye-fixations (produced from the DeepGazeIIE model [12]).

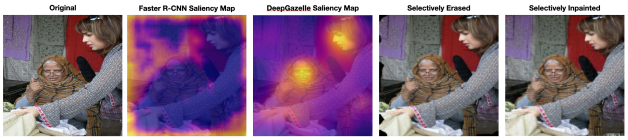

Figure 1. Example of selective erasing and selective inpainting.

While the MAE to some extent can reveal how similar the saliency maps are, we also calculate intersection over union (IoU) between the top 90% salient pixels of the generated saliency map and the top 90% salient pixels from the human attention map/predicted human eye-fixation. Calculating the IoU can help reveal whether the most salient region identified by the model and the humans align [2].

3.2. Data Augmentations

Data augmentation for object detection is more complex than that of image classification tasks because of the associated bounding boxes for each object. With this in mind, we designed three different data augmentation techniques: ***selective erasing***, ***selective inpainting***, and ***non-trivial transformations***. Below, we define and provide examples of each of these data augmentations.

Selective Erasing. The goal of selective erasing is to get rid of potential spurious patterns, patterns that the model has learned to associate with a label even though it does not represent that label. In order to augment images using selective erasing, we send the image through Faster R-CNN and use EigenCAM [14] to generate the saliency map from the layer 4 in the backbone. We then send the image through the DeepGazeIIE model [12] to generate the predicted eye-fixations map. We calculate the intersection over union (IoU) between the two saliency maps using a 90% threshold. If the IoU is below 0.1, meaning the two saliency maps are extremely different from one another, then we erase the top 2.5% salient pixels identified from the Faster R-CNN saliency map from the original image. We identified 6476 images that met this criteria. An example of this process and the outcomes from each step are shown in Figure 1. We chose the top 2.5% because these pixels would most likely make up the core regions of a potentially spurious region.

Selective Inpainting. The selective inpainting augmentation follows the same steps as selective erasing and then inpaints the erased image. To inpaint the top 2.5% salient pixels as denoted by Faster R-CNN, we send the selectively erased image and mask into an untrained neural network which learns the pixels by minimizing the loss function [17]. We used 4001 iterations with an untrained ResNet to inpaint the erased regions in each image. We augmented

Figure 2. Sample of some non-trivial augmentations used.

6476 images and replaced those images in the original PASCAL VOC 2012 to make up the final augmented dataset.

Non-trivial Transformations. To improve the model generalization, we apply the following augmentation techniques. In our work, we do experiments of bounding box geometric augmentation, color augmentation, and geometric augmentation. Each image in the dataset was augmented only once with a random augmentation. A subset of augmentations considered are shown in Figure 2. This was to ensure we had the same amount of data to fine-tune on as the other augmentation techniques.

3.3. Experiment Design

We gather a baseline to compare our three different data augmentations against. We fine-tuned Faster R-CNN on the PASCAL VOC 2012 training set and save the model to later evaluate it the our PASCAL VOC 2012 test set. During evaluation, we calculate the mean average precision (mAP) at IoU of 0.5. We also calculate the MAE and IoU between the saliency maps generated by the saved model and the predicted eye-fixations. We again calculate those metrics for the saliency maps generated by the model and the human attention masks.

For evaluating the impact of data augmentation, we created three different augmented PASCAL VOC 2012 training sets, one for each augmentation. Then, we separately fine-tuned the pre-trained Faster R-CNN on each augmented dataset. We do the same metric calculations as we did for the baseline model (mAP, IoU, and MAE).

For all fine-tuning, we used the following training parameters: 5 epochs, learning rate of 0.005, SGD optimizer with momentum of 0.9 and weight decay of 5e-4, and the StepLR scheduler with a step size of 2 and a gamma of 0.1.

4. Results

We present results from our empirical study and our main experiment which evaluates the impact of different data augmentation techniques. The empirical study was done to get a glimpse at how current saliency maps from state-of-the-art models compare to predicted and ground truth human attention. The main experiment extends the empirical study by evaluating the impact of different data augmentation techniques on the saliency maps.

4.1. Empirical Study

We compared the saliency maps from the models and the predicted eye-fixations as well as human attention masks for a subset of PASCAL VOC 2012 (898 images). This subset was determined based on the PASCAL VOC 2012 images that had a ground truth human attention map in the ML Interpretability Evaluation Benchmark. We calculated the mean absolute error (MAE) and intersection over union (IoU) for each model. The MAE (mean absolute error) is preferred to be close to 0; the IoU (intersection over union) is preferred to be close to 1. We calculate IoU using the top 90% salient regions.

We observed that Faster R-CNN with a ResNet50 backbone generated saliency maps are most similar to the predicted eye-fixations and the human attention masks in terms of MAE with values of 0.1700 and 0.1145. In terms of IoU, the SSD with a VGG backbone generated saliency maps are most similar to the predicted eye-fixations and human attention masks with values of 0.2474 and 0.3225.

A top preforming model is identified in terms of MAE because this metric is not variable based on a threshold like IoU. We selected the top performing model for our main experiment to focus on the impact of the augmentations for one model instead of comparing across different models. Since Faster R-CNN performed the best for MAE on PASCAL VOC 2012, we use this model in our main experiment.

4.2. Main Experiment

A pre-trained Faster R-CNN is fine-tuned on each data augmentation technique and then evaluated on the test images. We calculate mAP to understand the performance of each model and we also calculate MAE along with IoU between the saliency maps generated by the model and the predicted eye-fixations or human attention masks.

The Faster R-CNN generated saliency maps that were

Compared to Predicted Eye-Fixations			
Augmentation	mAP	MAE	IoU
Selective Erasing	0.754	0.1560	0.1878
Selective Inpainting	0.763	0.1552	0.1863
Non-trivial	0.781	0.1581	0.1762
No Augmentation	0.787	0.1575	0.1823

Table 1. Faster R-CNN fine-tuned on PASCAL VOC 2012. The mAPs reported are for IoU = 0.5.

Compared to Human Attention Masks			
Augmentation	mAP	MAE	IoU
Selective Erasing	0.754	0.1561	0.1878
Selective Inpainting	0.763	0.1572	0.1863
Non-trivial	0.781	0.1600	0.2676
No Augmentation	0.787	0.1583	0.2688

Table 2. Faster R-CNN fine-tuned on PASCAL VOC 2012. The mAPs reported are for IoU = 0.5

more similar to predicted eye-fixations in terms of MAE when using selective inpainting augmentation and in terms of IoU when using selective erasing. Table 1 shows that these augmentations impacted the mAP by at most 3%. When comparing the generated saliency maps to the human attention masks from the ML Interpretability Evaluation Benchmark [13], only selective erasing and selective inpainting improved the MAE as shown in Table 2.

5. Conclusion

Overall, we conduct two studies to understand how current object detection models compare to human attention and what techniques might improve them. We evaluate three novel data augmentation pipelines to see if they help saliency maps become more human centered without significantly sacrificing the accuracy. With at most 3% mAP difference, we observe that data augmentations that are derived from predicted human attention can improve the mean absolute error and intersection over union between the model saliency and predicted attention. In the experiment, selective erasing and selective inpainting augmentations only used the predicted eye fixations to create the augmented training dataset which could explain the null results shown in Table 2. Future works should create a separate augmented dataset using the human attention masks instead of predicted eye-fixations.

References

[1] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. Evaluating saliency map explanations for convolutional neural networks: A user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, page 275–285, New York, NY, USA, 2020. Association for Computing Machinery. 1

[2] Angie Boggust, Benjamin Hoover, Arvind Satyanarayan, and Hendrik Strobelt. Shared interest: Large-scale visual analysis of model behavior by measuring human-ai alignment. *CoRR*, abs/2107.09234, 2021. 2

[3] Aidan Boyd, Kevin W. Bowyer, and Adam Czajka. Human-aided saliency maps improve generalization of deep learning. *CoRR*, abs/2105.03492, 2021. 2

[4] Aidan Boyd, Patrick J. Tinsley, Kevin W. Bowyer, and Adam Czajka. Cyborg: Blending human saliency into the loss improves deep learning. *CoRR*, abs/2112.00686, 2021. 1

[5] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. Onboarding materials as cross-functional boundary objects for developing ai assistants. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA, 2021. Association for Computing Machinery. 1

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

[7] Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric T. Heim, Howie Choset, and Matthew E. Gaston. xbd: A dataset for assessing building damage from satellite imagery. *CoRR*, abs/1911.09296, 2019. 1

[8] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. Saliency benchmarking made easy: Separating models, maps and metrics. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, pages 798–814. Springer International Publishing. 2

[9] Qiuxia Lai, Salman Khan, Yongwei Nie, Hanqiu Sun, Jianbing Shen, and Ling Shao. Understanding more about human and machine attention in deep neural networks. *IEEE Transactions on Multimedia*, 23:2086–2099, 2021. 1

[10] Q. Vera Liao and Kush R. Varshney. Human-centered explainable AI (XAI): from algorithms to user experiences. *CoRR*, abs/2110.10790, 2021. 1

[11] Akis Linardos, Matthias Kümmerer, Ori Press, and Matthias Bethge. Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. *CoRR*, abs/2105.12441, 2021. 2

[12] Akis Linardos, Matthias Kummerer, 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 (ICCV)*, pages 12919–12928, October 2021. 2

[13] 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. 2, 4

[14] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. *CoRR*, abs/2008.00299, 2020. 2

[15] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. https://distill.pub/2017/feature-visualization. 1

[16] Gregory Plumb, Marco Túlio Ribeiro, and Ameet Talwalkar. Finding and fixing spurious patterns with explanations. *CoRR*, abs/2106.02112, 2021. 1

[17] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Deep image prior. *CoRR*, abs/1711.10925, 2017. 2

[18] B. Zhou, A. Khosla, Lapedriza. A., A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. *CVPR*, 2016. 1