Designing and Evaluating Universal Attack on Explainable AI

Chaewon Park (cp3227), Noah McDermott (ntm2128)

1. Abstract

Explainable AI has become an important and growing field because as black-box models are introduced in applications that have large social impact or put lives on the line, there is an increasing demand to understand why they arrive at the decisions they make. In particular, in computer vision, heatmaps have been used to demonstrate which areas of each image contribute to the decisions made. As with neural networks themselves, heatmap explanations have been shown to be vulnerable to adversarial attacks, which are slight changes in the input images that have the potential to completely change explanations.

We intend to be the first to develop a universal attack for heatmap explanations; a small perturbation that gives a bad explanation for any image, while still retaining the proper classification. Similar universal attacks have been developed for image classification systems. Our method focuses on maximizing the distance between the perturbed images and the original ones at the layer used to generate the explanation, while minimizing the distance at the final layer which outputs a classification. We used a series of distance metrics including L2, Dice Loss, and center-of-mass distance, and measured distances between both the activations of the neural networks and the heatmaps themselves. We achieved a success rate of 0.37-0.43 on the ResNet-18 model trained on the CIFAR-10 dataset, with GradCAM being used to generate the explanations.

2. Introduction & Related work

What is Explainable AI?

As Neural Networks are being widely applied in many fields, there has been a growing interest in explaining how they make decisions. Conventional neural networks were considered to be ‘black box’, meaning they only output a result and do not explain their course of prediction or rationale. This characteristic has made it very difficult to debug models if a prediction seems wrong and for human users to have trust in the model’s performance. Hence, there has been a new line of research dubbed ‘Explainable AI’, conducted on making the prediction process of a neural network to be more transparent, trustworthy, and ‘white box’ - by providing explanation to why a model made certain decisions. Explanations can vary in form, ranging from leveraging Natural Language Processing techniques to generate sentences or phrases as explanations, providing human-comprehensible numerical metrics that reflect the importance weights of input features, to generating heatmaps, also known as attention maps, that highlight areas in an image that contributed the most to the decision of the model.

Significance of Explainable AI

Explainable AI has consequently been implemented in domains that are safety-critical or requires a high level of social acceptance, such as medicine, autonomous vehicles, and finance. For instance, in hospitals, more and more doctors are utilizing AI to get second opinions on patient's lab results or scans. With Explainable AI, doctors can be provided with concrete explanations as to why the AI model made certain diagnoses and gain insight that otherwise might have been overlooked. This ensures that doctors make less mis-interpretations of the patients' illness, and also as these explanations are very intuitive and human comprehensible, it opens doors for non-medically trained people like the patients themselves to understand the AI model's diagnosis and have a strong sense of acceptance and trust in the medical service they receive. In the finance industry, Explainable AI can be leveraged in tracking credit card fraud and providing human comprehensible explanations to the credit card company so that they can take appropriate action, or for banks to provide mandatory explanations, which is required by law, as to why their internal assessment system rejected certain clients who requested credit loans. For autonomy, Explainable AI can be used to debug detection errors in vision systems and fix them before an actual accident happens, and also passengers can have a higher level of trust and sense of safety with the aid of these explanations.

Problems and Vulnerability of Explainable AI

However, it has been shown through various research that explanations can be manipulated by iteratively applying visually hardly perceptible perturbations to either the input image or model weights. This attack results in passive and active fooling that causes 'fair washing'. Passive fooling is an occurrence where visual explanations do not convey any meaningful information, such as heat maps that are formed in a noisy manner and not on a particular object in an image. On the other hand, active fooling is an occurrence where the adversary intentionally creates a wrong explanation, such as forming a misleading heatmap on a particular wrong object. In these cases, the classification prediction stays unmodified and correct but the adversarial attack messes with the visual explanation. This lack of robustness is especially problematic in applications that are safety-critical and requires a high sense of social acceptance and trust, which happens to be the areas where Explainable AI has most significance in.

The paper 'Fooling Neural Network Interpretations via Adversarial Model Manipulation' (NeurIPS 2019) tries to change explanations without impacting the model accuracy, by fine tuning the model weights. Our approach differs in that we don't change the model's weight, but is similar in that we attempt to maintain the model accuracy as is. The authors optimize a loss function that has a penalty term that directly incorporates the interpretation results. The results have a Fooling Success Rate (FSR) of 80-90%, with a 'success' defined as a distance metric (Spearman rank correlation) above some pre-chosen threshold.

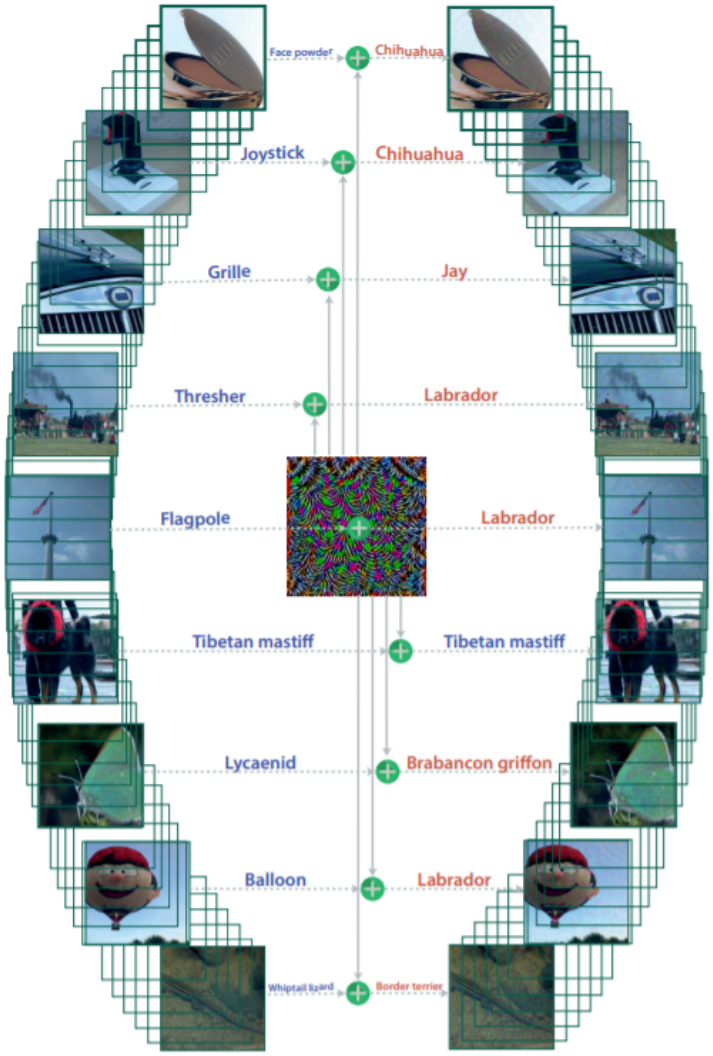
The paper 'Explanations can be Manipulated and Geometry is to Blame' (NeurIPS 2019) attempts to manipulate an input image's visual explanation to a target image's visual explanation and establish theoretically that this vulnerable phenomenon is related to certain geometrical properties of neural networks. This approach is different from ours in that this

method needs a target explanation to approximate to. The results are measured as a distance metric from the target explanation, and they achieve a very small distance.

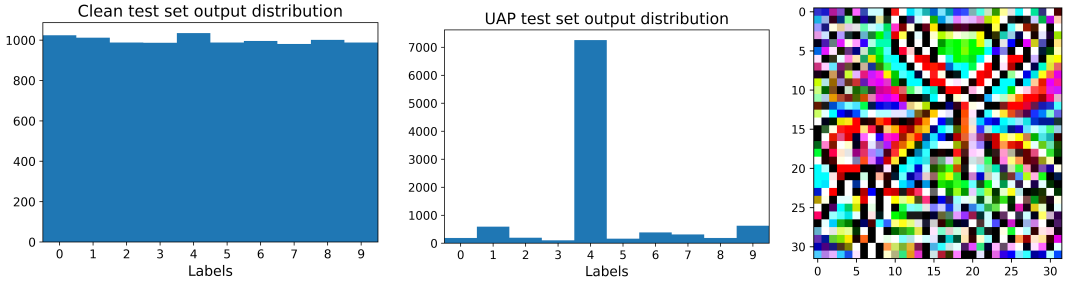
Universal Attack

Previous results have shown that neural networks are vulnerable to adversarial attacks consisting of slight changes to the input that causes the model to misclassify or otherwise malfunction. Typically, these attacks focus on a single image, but in 2017, the existence of universal adversarial attacks was proven, a single perturbation that could fool any image. (Moosavi-Dezfooli et. al.)

The original universal attack was a simple process. The first step is initializing a perturbation. Then, for each image in the training set, we compute the minimum change needed to get the network to misclassify, add it to the perturbation and scale. Hence, the perturbation is generated from combining a series of many successful attacks. To get the network to misclassify, the attack method known as DeepFool was used. (Moosavi-Dezfooli et. al.)



While this method delivers promising results, due to the fact that it uses a state-of-the-art method to attack every single image, it can be impractical if a large number of images is needed. In order to generate enough universal adversarial perturbations to perform adversarial training, a new method was developed to streamline the process, called Stochastic Projected Gradient Descent (sPGD). The overall method is similar, but instead of attacking and causing the network to misclassify, it takes a single step in the direction opposite of the gradient for each image, over an entire batch. Since it takes just one step instead of a single attack for each image, it is much more practical for generating many examples. A targeted attack with sPGD is shown below. (Shafahi et. al.)



Why it is hard or unsolved

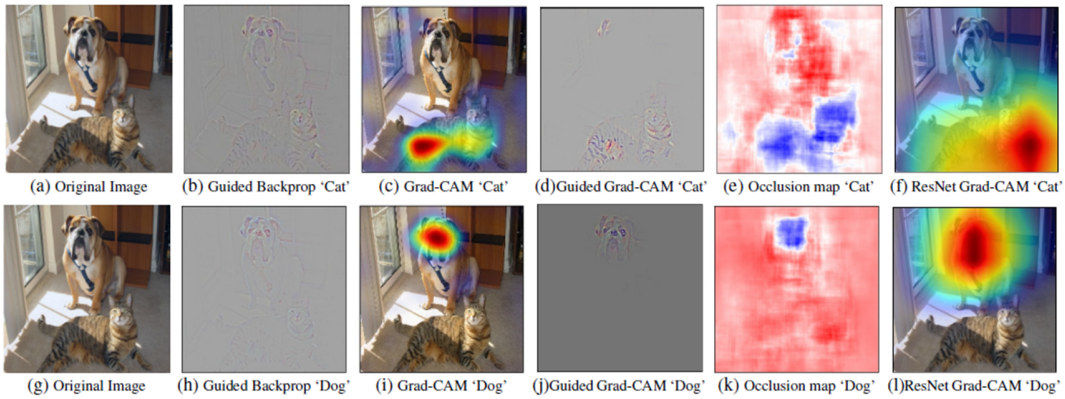
This topic has yet to be explored because Explainable AI itself is a relatively new research field, and most research on adversarial attack has historically been focused on messing with the model prediction, not the model explanation. Related works on adversarial attack on visual explanation have been slowly emerging since 2018, but none of them have attempted to find a single universal attack pattern in respect to the visual explanation.

Discovering the universal pattern for attacking visual explanation can reveal a critical security vulnerability, as the adversary can easily perturb the model using the same pattern with a high success rate, without having to compute and change the pattern for different classes or input every time. Also, this technique doesn't modify the model weights, so even if the adversary doesn't know the specific structure of the model (black box scenario), the attack can still occur.

Our Approach

We experimented with a classic visually explainable method for image classification task, GradCAM (Gradient-based CAM), on a ResNet18 backbone model with the CIFAR10 dataset. As can be derived from the equations below, GradCAM uses the gradient information flowing into the last convolutional layer of the CNN to assign importance values (α_k^c) to each neuron for a particular decision of interest. In other words, α_k^c captures the 'importance' of feature map k for a target class c. The final attention map is generated as a weighted combination of forward activation maps, followed by ReLU, which highlights the areas of the image that contributed most to the classifier's prediction.

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}} \qquad L^c_{\text{Grad-CAM}} = ReLU\underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}$$



(a) Original Image (b) Guided Backprop 'Cat' (c) Grad-CAM 'Cat' (d)Guided Grad-CAM 'Cat' (e) Occlusion map 'Cat' (f) ResNet Grad-CAM 'Cat'

(g) Original Image (h) Guided Backprop 'Dog' (i) Grad-CAM 'Dog' (j)Guided Grad-CAM 'Dog' (k) Occlusion map 'Dog' (l)ResNet Grad-CAM 'Dog'

For the attack, we leveraged the sPGD algorithm to attempt to generate a single perturbation that maliciously makes bad (i.e. uninformative, misleading) explanations when added to input images of any class, while remaining visually similar to the original. There were several reasons as to why this was chosen. First of all, it attacks in the direction of the gradient, instead of trying to move outside the class boundary like the original attack. When generating an attack on an explanation, there are no analogous boundaries to determine whether an explanation is good or bad. Second of all, the gradient descent attack is much faster, due to taking one step at a time, and since in some variants of the attack, an explanation must be generated every step, speed is a high priority. (Shafahi et. al.)

The loss function had to do two things: create a very different heatmap and keep the classification output the same. To maintain the same classification output, we used cross-entropy loss to ensure as many examples as possible were predicted correctly. We used several different pairing loss implementations to measure the distance between GradCAM explanations on the original and perturbed inputs. The GradCAM explanations were generated for both the perturbed and original images, and then a distance metric was taken, and this was repeated to optimize the perturbation.

Since evaluating the heatmap for every image used in training was time-consuming and impractical for any kind of adversarial training, we decided to simplify the process. Instead of backpropagating through the network and using both the gradients and the activations to compute a heat map, we only used the activations, which could be obtained quickly from the forward pass. We made the general assumption that activation functions that are very different would result in heatmaps that were very different. This was inspired by Adversarial Logit Pairing,

which used a similar distance metric between the activations, but in this case we used the layer in which the heatmaps were based off of, usually the last convolutional layer. (Kannan et. al.)

These loss functions were as follows:

1. L2 Norm: L2 norm is simply the euclidean distance between the original and perturbed heatmaps. A high L2 norm implies a large difference between the two heatmaps.

$$L_{L2} = \sqrt{\sum_{i=1}^{n}(x_i - \hat{x_i})^2}$$

2. Dice Loss: Dice loss measures the size of the intersection between the perturbed and original heatmap over the total size. A low dice loss means that there is very little overlap between the original heatmap and the perturbed one, implying a big difference between them. Hence, our goal is to minimize the dice loss.

$$L_d = \frac{2|x \cap \hat{x}|}{|x| + |\hat{x}|}$$

3. Center-of-mass distance: The center of mass of both heatmaps was measured, and the L2 distance between the two was taken. (Heo, et. al.)

Due to the complexity of optimizing center-of-mass distance over the generated heatmaps, center-of-mass was only used with the raw activations. This approach differs from the non-universal attacks because the stochastic gradient descent algorithm moves one step in the gradient in each image in the batch, while a non-universal attack would not generalize and focus on just one image.

We found that none of the metrics used in the non-universal attacks or the explainable models themselves were sufficient to evaluate success. The Fooling Success Rate (FSR) required a threshold function specific to the loss function being evaluated, and we felt that setting the cutoff for what is considered a success could introduce bias. (Heo et. al.) GradCAM and other explainable models used Amazon Mechanical Turk to have humans evaluate the validity of explanations, which was impractical for the scope of this project. (Selvaraju et. al.)

We chose the attack success metric as (1-dice loss), where the dice loss is the average dice loss over many inputs, between the original attention map and the perturbed attention map. The dice loss is an intersection over union metric and therefore, if the attack was performed successfully and ideally, the two heatmaps would not overlap and the dice loss would be closer

to 0. On the other hand, if the attack failed and the two attention maps overlap on many areas, the dice loss would be closer to 1. Our average attack success rate was around 0.41.

3. Description of what you did/built

When implementing the pairing loss, we extensively experimented with five different methods. The first three methods use the raw activations from the intermediate layers, and the other two methods use the weighted activations which are also referred to as attention maps. We devised different distance metrics to see which methods optimize the best. Also, we fine tuned the size of the perturbation(EPS) so that all model instances have a comparable accuracy of around 75%.

Input	Distance Metric	EPS	Model Accuracy	Attack Success Metric (1-dice)
Activation	L2 Norm	3.5/255	74.5%	0.41
Activation	Dice Loss	7/255	75%	0.44
Activation	Center Mass	9/255	73%	0.43
Attention Map	L2 Norm	12/255	73.4%	0.42
Attention Map	Dice Loss	12/255	75%	0.37

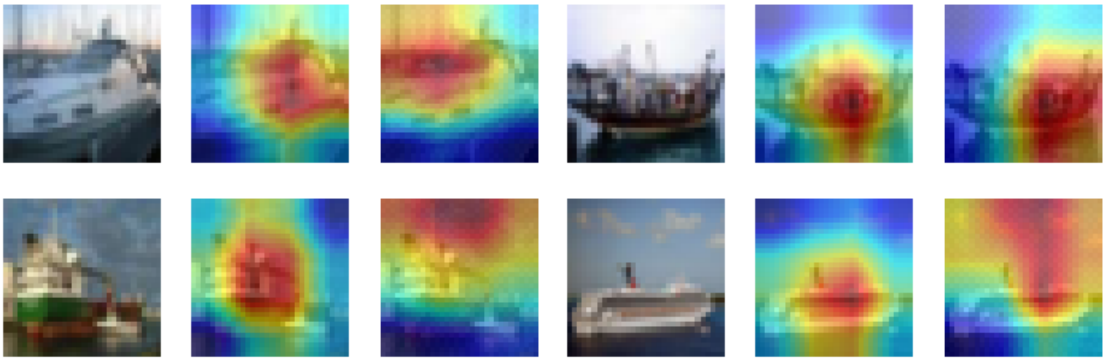
To provide description on the distance metrics, the center mass finds the respective center points of the two activation blobs and try to maximize the distance between them so that the bad visual explanation would be as different and far as possible from the true explanation. When applying L2 Norm and dice loss to the attention map, the attention maps which have the RGB value range of [0:255] were hard-thresholded with 150 to make the comparison easier.

4. Results

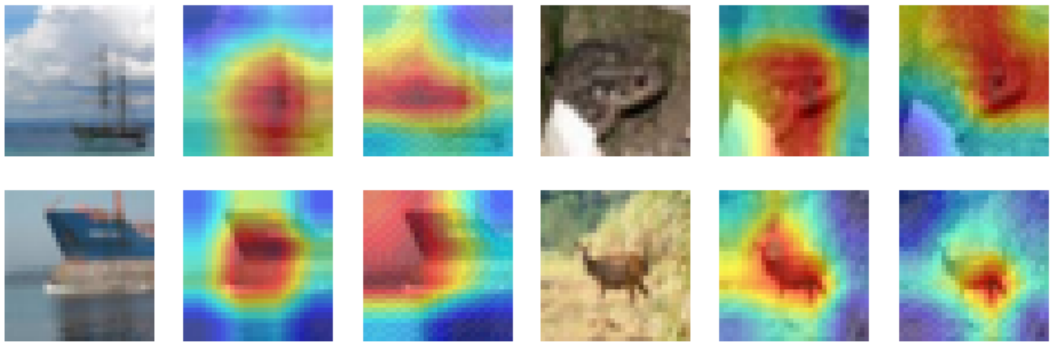
One lesson learned is that visual explanation is very hard to assess quantitatively using a metric. This is an on-going bottleneck issue present in the research field of explainable AI. Although the soundness and quality of attention maps can only be qualitatively and visually measured, we did our best to use the dice loss as a proxy metric to evaluate the quality. However, sometimes the dice loss doesn't reflect the quality realistically so we also relied heavily on many manual evaluations on the attention map output for better analysis. Below are the sample outputs of each loss function. The table in Section 3 shows the attack success rate and experiment settings of each model.

We only looked at images where both the clean model and perturbed model correctly predict the

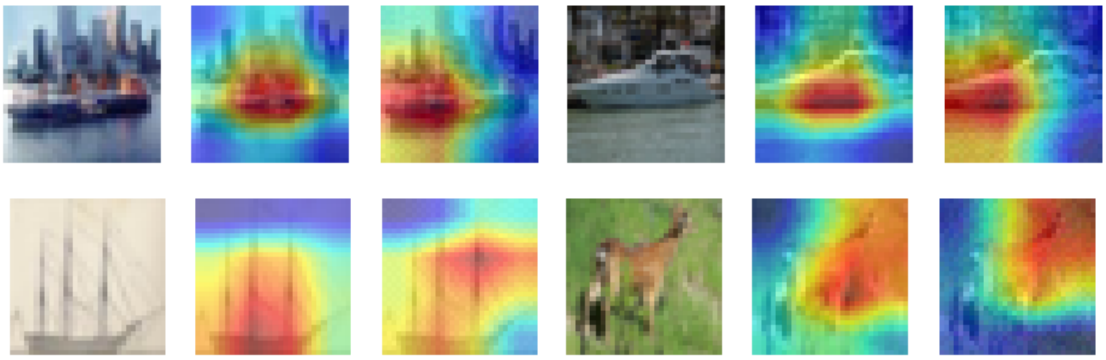
label as ground truth label. The first column shows the original input image, the second column shows the original attention map, and the third column shows the perturbed attention map. Overall, it is clearly illustrated that the perturbation causes the attention map to shift to wrong or irrelevant areas of the image that provides less information and intuition as to why the model made a certain class label prediction. For instance, for images displaying a ship, the original attention maps usually highlight the full body of the ship. However, a perturbed attention map shows that the attention has shifted to the sky or edges of the boat, even when it has correctly classified the image as a 'ship'.



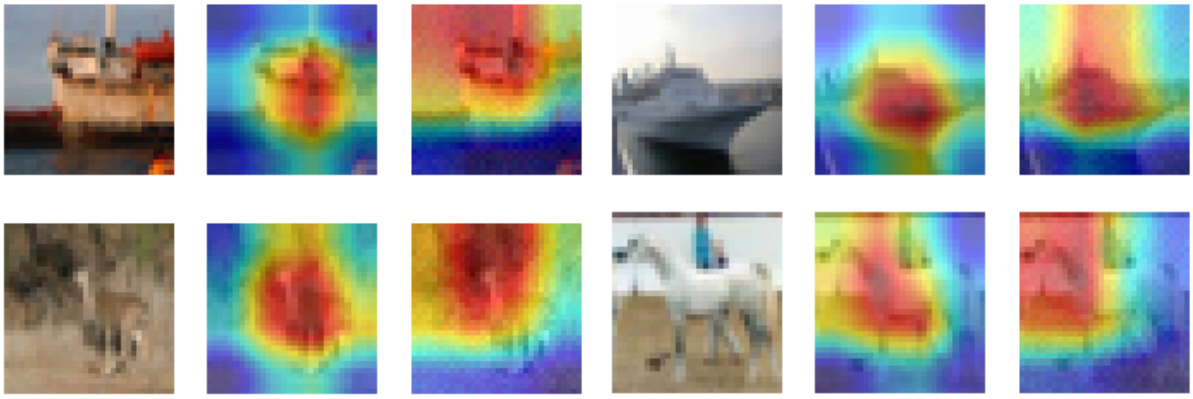
Raw Activations - L2 Norm



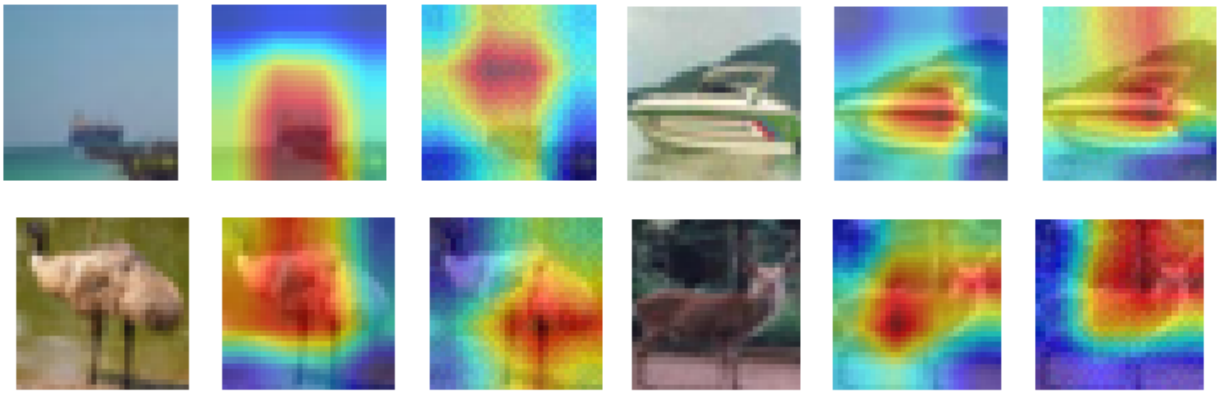
Raw Activations - Dice Loss



Raw Activations - Center Mass



Attention Maps - L2 Norm



Attention Maps - Dice Loss

Since the attacks performed similarly on both the raw activations and the heatmaps, we can conclude that we were correct in our assumption that activations are a good proxy for attention maps, and that a different activation usually leads to a different heatmap.

5. Conclusions

One of the most significant takeaways from our project is the fact that a universal attack on visual explanation is possible, and we were the first ones to design and test out this idea. However, the success rate was not as high as initially expected. We believe this is due to several factors.

First of all, our loss function has two competing loss terms with opposite goals, and therefore it is not easy to balance and optimize. From papers, universal attack on classification achieves around 80% success rate, and this is because its loss function is simply maximizing the Cross Entropy to make the classifier misclassify as much as possible. Optimizing such a single term is definitely a more trivial task than how our delicate and multi-composite loss term is set up. The second reason is that unlike attack on classification labels, attack on visual explanation has to change and handle so much information- consisting of an enormous amount of pixels corresponding to the activation maps and its corresponding weights, to varying pixel intensities and correlations. Also, we need to generate a universal pattern that generalizes for an enormous amount of different input images. Therefore, we hypothesize that it is possible that a single universal pattern may be too weak a tool that does not have enough capacity to handle and perform a universal attack on visual explanation. Therefore, the universal pattern does not generalize over diverse input images as much as we expected it to.

Furthermore, universal attack and the generation of a universal pattern relies very heavily on the training data. A very small change in the training dataset changes the universal pattern entirely. This led us to think that since universal pattern extraction is so sensitive to the characteristics of the training data, it could be possible that the intrinsic qualities of CIFAR10 could have impacted the success rate of the attacks. CIFAR10 in general has very low resolution and is 32 by 32 in size. Another glaring characteristic is that all the images have a single instance object placed right in the center of the image. Therefore it could be very challenging for the attention maps of the GradCAM to not overlook or miss them, and therefore is not able to learn a universal pattern that generates the wrong attention map.

Based on these findings and insights, our proposed next steps is to try further experimentations on a different dataset that is bigger in size, scale, resolution, and has more rich diversity with multiple instances and positions, like ImageNet. In addition, it would also be an interesting problem to compare the attack success rates with Class-Discriminative Universal Attacks. Class Discriminative Universal attacks are ‘less’ universal because rather than attacking all existing classes, it makes its attacks more focused on selective classes. This is actually considered a more realistic and threatening attack because in reality, attackers want their attack to go unnoticed, but with universal attack, it is easier for people to notice since the attack is occurring to every class. Class-discriminative universal attack opens the door to a more stealthy attack that could induce serious security problems.

References

Zhang, Chaoning, et al. "A survey on universal adversarial attack." *arXiv preprint arXiv:2103.01498* (2021).

Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

Shafahi, Ali, et al. "Universal adversarial training." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. No. 04. 2020.

Mummadi, Chaithanya Kumar, Thomas Brox, and Jan Hendrik Metzen. "Defending against universal perturbations with shared adversarial training." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.

Deng, Yingpeng, and Lina J. Karam. "Universal adversarial attack via enhanced projected gradient descent." *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020.

Heo, Juyeon, Sunghwan Joo, and Taesup Moon. "Fooling neural network interpretations via adversarial model manipulation." Advances in Neural Information Processing Systems 32 (2019): 2925-2936.

Dombrowski, Ann-Kathrin, et al. "Explanations can be manipulated and geometry is to blame." *arXiv preprint arXiv:1906.07983* (2019).

Kannan, Harini, Alexey Kurakin, and Ian Goodfellow. "Adversarial logit pairing." *arXiv preprint arXiv:1803.06373* (2018).

Moosavi-Dezfooli, Seyed-Mohsen, et al. "Universal adversarial perturbations." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.