ADVERSARIAL EXAMPLES FOR IMAGE CROPPING IN SOCIAL MEDIA

Masatomo Yoshida[⋆], *Masahiro Okuda*[†]

Doshisha University, Kyoto, Japan
[⋆]masatomo.yoshida@mm.doshisha.ac.jp, [†]masokuda@mail.doshisha.ac.jp

ABSTRACT

It is well known that constructing adversarial examples is important to make a machine learning system more stable by pointing out its vulnerabilities. We present an approach to produce the adversarial examples to image cropping systems. In this method, we propose a method to calculate repeated perturbations based on the gradient information calculated using a model similar to the cropping system provided by a social network service, and give the cropping model an attack to change regions that should be cropped. We measure the amount of shift in the peak value of the saliency map and quantitatively verify the effectiveness of the system by this amount. We demonstrate that the proposed method outperforms baseline methods.

***Index Terms*—** image cropping, object detection, adversarial examples, saliency map, Twitter

1. INTRODUCTION

There is a lot of demand for cropping images due to lack of space to display or to emphasize regions of interest, and Machine Learning (ML) models are now commonly used for automatic cropping. Twitter[1], for example, has announced that when displaying user posts, it uses ML models to create thumbnails by cropping the posted images into appropriate display areas. Netflix also uses an ML model to extract a meaningful scene from a movie to make thumbnails[2].

On the other hand, the study of adversarial examples for ML models has been actively conducted [1–5]. These researches are not just attacks that exploit the vulnerabilities of ML systems but also contribute to system stability and improvement on generalization performance.

In this paper, we develop a method to generate adversarial examples that shift the cropped area of images (Figure 1) by applying gradient-based attacks to the model, which predicts a saliency map. While previous studies [1–4, 6] focused mainly on classifiers or detectors, the goal of our method is to perturb input images and give a cropping model an attack to change the regions that should be cropped. The proposed method levers a technique called FGSM, introduces a fundamental approach to the novel type of ML model, and realizes effective attacks against the objective model. The contributions of this paper are as follows:

1. We create a Deep Neural Network model based on an image cropping system actually used in a social networking service, and propose an iterative perturbation generation algorithm using the model.
2. We also introduce a new method to evaluate the image cropping quantitatively and verify the validity of adversarial attacks to show the effectiveness of the proposed method.

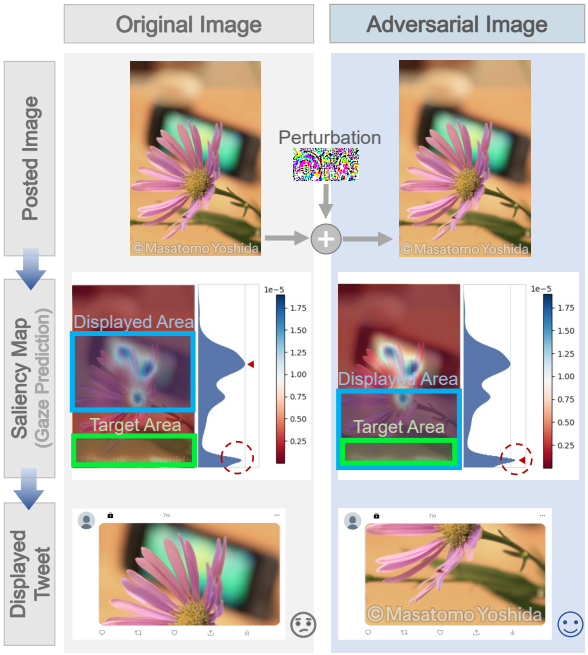

Fig. 1. Overview of the proposed method[3]. In the original image (left), the copyright notice on the bottom is not displayed. However, the proposed method (right) adds perturbations that lead cropping functions to show the desirable area.

[1]https://twitter.com/
[2]https://netflixtechblog.com/a442f163af6

[3]The image is illustration purpose only (not included in the dataset).

2. RELATED WORKS

2.1. Adversarial Examples

Adversarial examples were first generated by Szegedy et al. [1] by adding small perturbations to lead state-of-the-art DNN models to misclassification. Goodfellow et al. [2] proposed a fundamental method called FGSM, which is based on the gradient of the images. Typical applications of adversarial examples are [3, 6]. Recent works on image cropping or scene extraction [7, 8] focus on reinforcement learning.

Papernot et al. [9] introduced transferability attacks that use similar models' weights and are available to black-box models. Ghorbani et al. [10] showed small systematic perturbations to the model that create feature importance maps. This paper is closely related to these two papers.

To the best of our knowledge, there is very little research on attacks on the cropping models, while much of the previous researches on adversarial examples have been on classification and detection.

2.2. Saliency Map

Many methods including ours utilize the "saliency map" to generate adversarial attacks, which mainly has two roles. One is the extent of gaze concentration measured by eye-tracking or predicted (calculated) gaze concentration. Many models including Itti et al. [11] and Ardizzone et al. [12] are used in image processing, segmentation, and object detection. Twitter, mentioned above, published in their blog that they use DNN models of gaze prediction to crop images uploaded by users when displaying posts[4]. This model is based on Kummerer et al.'s model called DeepGaze II [13].

Another role of saliency map is to express feature importance by the map representing which feature affects the model's output. Ghorbani et al. [10] referred to the term as this meaning. Most of the previous studies other than Deep Gaze II made the saliency map (of feature importance) from the value of the hidden layer (not the output layer).

3. PROPOSED METHOD

3.1. Adversarial Attack

Our method adopts the model used by Twitter [14] as an example, which is a typical neural network model for the image cropping. Our approach creates perturbations based on the gradients. The system mainly takes as input an original image and the area specified by a bounding box that an attacker wants to display instead of an area that should be displayed. We repeatedly calculate the perturbations using the gradient of the model. The proposed method is a modification of the conventional method proposed for image classification in the image cropping task, and features repeated perturbations to increase the effectiveness of the method.

[4]https://blog.twitter.com/engineering/en_us/topics/infrastructure/2018/Smart-Auto-Cropping-of-Images.html

Algorithm 1 Perturbations Created by the Proposed Method

Input: Original Image $\boldsymbol{x}$, Target area (to be displayed) $\boldsymbol{y}$, parameter to adjust perturbation size α, # of iteration N
Output: Adversarial Image $\boldsymbol{x'}$

```
x' = x
for k ← 1 to N do
    x_p = x'
    η' = ∇_x J(θ, x', y)   // Calc. perturbation
    η = α · η'/||η||_2   // Adjust size of perturbation
    x' = x_p + η   // Add perturbation
end for
return x'
```

The proposed method combines the approaches that control the interpretation of the model mentioned in Section 2.1 and the gaze prediction model based on the saliency map discussed in Section 2.2. Our objective is to add perturbation to an image to shift the salient area and vary the area to be cropped by the target model.

As with other conventional methods such as [15], this method uses a gradient-based approach. The adversarial attack using FGSM is known to be less perceptible to perturbations, but it has been pointed out in [16] that it is less effective for models that include blurring operations, such as the one targeted by our method, and for this reason, we use a method similar to FGM instead of FGSM.

In our method, the perturbation $\boldsymbol{\eta}$ is calculated as the following equation:

$$\boldsymbol{\eta} = \epsilon \cdot \nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y}), \tag{1}$$

where $\boldsymbol{\theta}$ are the parameters of a model, $\boldsymbol{x}$ is the input image to the model, ϵ is a parameter to adjust the perturbation size, $J(\cdot)$ is the cost of training. In this paper, $\boldsymbol{y}$ represents the target area for the input image. The target area here is the area that the user may want to display. Our method applies an operation in (1) several times with a smaller perturbation size. In this paper, perturbations are created with 3 iterations, and ϵ is calculated to meet the perturbation size specified in the experiment. Processes to create perturbations are shown in Algorithm 1.

As shown in (1), the proposed method needs the gradient information of the objective model, but it was unavailable to get the weights in the desirable form through the model. Thus, we infixed the output of the hidden layer ahead of the readout network and re-trained the model.

The proposed method in this paper is different from previous studies in two perspectives:

1. The proposed method shifts the gaze prediction, which is the output of the final layer (output layer). In contrast, previous studies mainly used the feature importance, an output of the hidden layer (not the output layer).

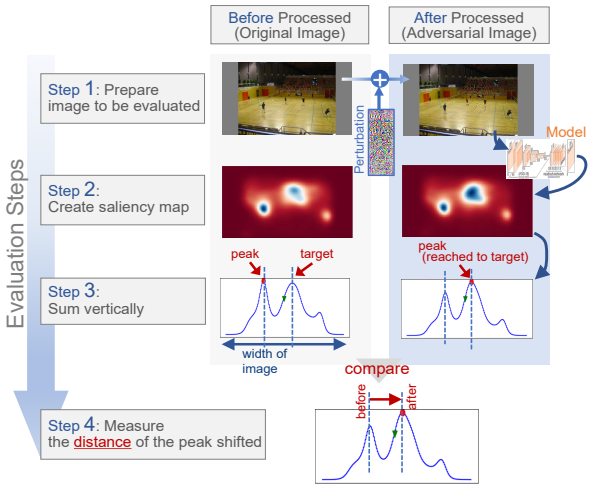

Fig. 2. Procedure of the evaluation.

2. The model in the proposed method contains a blur layer ahead of the output layer. Brama et al. [16] showed if the model contains strong blur effects before the input layer, the performance declined with adversarial examples. As most previous studies use too small perturbations, they can be less effective when the model has blurring effects.

3.2. Evaluation method

The saliency map of the gaze prediction gives us meaningful information which we can use for image cropping. We use subjectively evaluate the validity of the algorithm using the change in the saliency map before and after the attack as follows. Pixels of the saliency map are added vertically to form a one-dimensional sequence, and we refer to its maximum value as a "peak." Another is the point of the mass center projected on the x-axis. Because it is harder to move the peak than the mass center, the former ("peak") is a reasonable value to be examined. In each image, we assign the target to the second highest local maximum of the saliency map. We evaluate the validity of the attack by the amount of the peak movement in the saliency map.

Figure 2 shows the procedure of evaluation. The first step is to create a saliency map for the images before and after the insertion of the perturbation via the model. We measure how close the peak of the saliency map is to the target, and evaluate the validity of the method by the amount of movement.

4. EXPERIMENT

4.1. Setting and Dataset

The images from CAT2000 [17] are used as the dataset. This dataset contains original saliency maps of gaze concentrations made from their experiment with humans. Since the proposed method aims to shift the peak of images that have multiple local maxima, images with single or more than four local maxima are excluded from the dataset. Other major restrictions applied to exclude images that are inappropriate for the experiments are as follows (if an image meets at least one of these restrictions, it is excluded). (1) Vertical photos[5]: photos in which the filled pixels are more than 25% (which corresponds to the aspect ratio of 4:3). 0% corresponds to 16:9. We should note that this method is applied only to horizontal images in the experiments, but it can be applied to images of any aspect ratio. (2) distance of the x-axis between the peak and the target is more than 15%, compared to the width of the images. (3) distance of the y-axis (height) between the summed value of the peak and that of the target is more than 40%, compared to the summed value of the peak. After all, we used 335 images in this experiment and resized them to 640 pixels in width. Similar to FGM, our method uses Grad-CAM [18] to reduce the area where the perturbation is applied. Grad-CAM is a visualization method of important regions proposed by Selvaraju et al. and can reduce unimportant areas. The output of Grad-CAM was normalized to the scale of 0 to 1 and multiplied with the output of our method. In other words, our output is masked by the importance map of Grad-CAM. Because the gradient is re-calculated in every step in our method, ϵ is 1.77 times smaller than that for FGM.

Table 1. Distance of the peak shifted, compared to the distance to the targets (Averaged).

Perturbation size (L2 norm)	10	20	30
Gaussian Noise only	4.2%	9.6%	12.5%
Grad-CAM x Gaussian Noise	5.5%	9.5%	11.2%
Grad-CAM x Gradient (FGM)	40.9%	61.9%	69.2%
Proposed Method	**50.0%**	**65.8%**	**74.4%**

The baseline methods are three-fold. One method uses Gaussian noise only, and the second method uses Gaussian noise and Grad-CAM simultaneously. The third one is the gradient-based method, where FGM is applied to saliency transfer. When applying Gaussian noise, the area to be applied was limited to around the target point. We set the width of the area to 100 pixels (which corresponds to 15.6%, compared to the image width) and set the height to the same height as the image.

As an evaluation metric, the distance of the peaks shifted by the perturbation, divided by the distance to the target, was used. To get an accurate evaluation, we implemented the measurement with the saliency maps generated through the original Deep Gaze II model [13], which is almost the same as Twitter's cropping system.

4.2. Results

Results are shown in Table 1, 2, Figure 3, 4 and 5. Table 1 shows the distance of the peak shifted. The percentage

[5]Note the difference with Figure 1 (vertical photo). As 67% of the dataset are horizontal photos, we used horizontal photos in the experiment.

Table 2. Distance of the peak shifted, compared to the width of each image (Averaged).

Perturbation size (L2 norm)	10	20	30
Gaussian Noise only	1.2%	2.7%	3.6%
Grad-CAM x Gaussian Noise	1.6%	2.7%	3.2%
Grad-CAM x Gradient (FGM)	10.9%	16.2%	18.6%
Proposed Method	**12.9%**	**17.2%**	**19.0%**

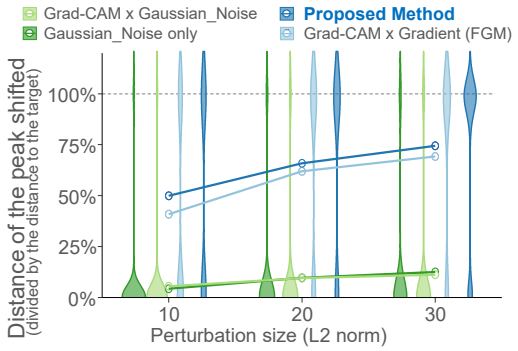

Fig. 3. Distribution of the distance of the peak shifted.

in the table shows the distance normalized by the distance between the original peak position and the target. The higher the value, the closer to the target and the larger the change in saliency area. The proposed method performed much better than the baseline in all perturbation sizes (Table 1). Gaussian Noise with Grad-CAM outperformed Gaussian Noise only in a small perturbation. In contrast, Gaussian noise without Grad-CAM outperformed in a large perturbation. The gradient-based baseline method (FGM) outperformed the other two baseline methods in all perturbation sizes. For reference, we also implemented the evaluation by replacing the divisor (the distance to the target) with the width of each image (Table 2) and obtained similar results. Figure 3 shows the distribution of the distance of the peak shifted. We can also get remarkable insight from the distribution, that is, the distance in each image tends to fall near 0% or 100%, not between them.

Experimental results are shown in Figure 4. Figure 5 shows the comparison of the proposed and baseline methods in each step. As shown in the image, it contains specific patterns rather than the baseline method, but they are still less perceivable.

5. CONCLUSION AND FUTURE WORK

This paper introduces a framework to create adversarial examples against the gaze prediction model used in web services, focused on the ML model in Twitter. Key points of the proposed approach are as follows:

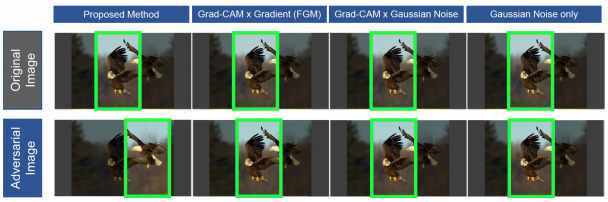

Fig. 4. Experimental Results: Area surrounded by a square indicates the cropped area.: (upper row) original image, (bottom row) adversarial examples created by several methods.

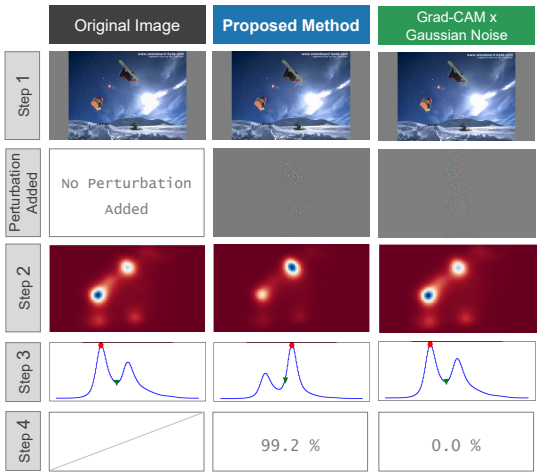

Fig. 5. Comparison among original, proposed, and baseline methods. The step numbers (Step n) correspond to the numbers in Figure 2.

1. We leveraged existing adversarial examples and introduced a fundamental approach to create adversarial examples against the saliency detection model (to meet the needs).
2. The proposed method proved to be effective to the model with blur layer (which was pointed out to be a harder condition).

Based on the proposed method, meeting the existing needs to shift the cropped area, reducing the potential legal risk, and improving more stable systems are expected. However, the proposed methods have some limitations. For example, the model used in the proposed method includes the blur layer, but the layer is not in exactly the same position. This suggests that enough size of perturbations is not the same level as the previous studies. In other words, it indicates this model requires larger perturbations than the previous models. In addition, the proposed method adds the perturbations on the target area, but it can be added on the other area, such as edges in future work. The creation of harder-to-perceive perturbations under this limitation and the effectiveness on larger datasets are parts of future work.

6. REFERENCES

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, "Intriguing properties of neural networks," in *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 2014.

[2] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[3] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu, "DolphinAttack: Inaudible voice commands," in *Proceedings of the ACM Conference on Computer and Communications Security*, New York, NY, USA, 2017, CCS '17, pp. 103–117, Association for Computing Machinery.

[4] Pin Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho Jui Hsieh, "ZOO: Zeroth order optimization based black-box atacks to deep neural networks without training substitute models," in *AISec 2017 - Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, co-located with CCS 2017*, New York, NY, USA, 2017, AISec '17, pp. 15–26, Association for Computing Machinery.

[5] Joern-Henrik Jacobsen, Jens Behrmann, Richard Zemel, and Matthias Bethge, "Excessive Invariance Causes Adversarial Vulnerability," in *International Conference on Learning Representations*, 2019.

[6] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song, "Robust Physical-World Attacks on Deep Learning Visual Classification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1625–1634.

[7] Debang Li, Huikai Wu, Junge Zhang, and Kaiqi Huang, "Fast a3rl: Aesthetics-aware adversarial reinforcement learning for image cropping," *IEEE Transactions on Image Processing*, vol. 28, no. 10, pp. 5105–5120, 2019.

[8] Evlampios Apostolidis, Eleni Adamantidou, Vasileios Mezaris, and Ioannis Patras, "Combining adversarial and reinforcement learning for video thumbnail selection," New York, NY, USA, 2021, ICMR '21, p. 1–9, Association for Computing Machinery.

[9] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow, "Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples," *arXiv preprint arXiv:1605.07277*, 2016.

[10] Amirata Ghorbani, Abubakar Abid, and James Zou, "Interpretation of neural networks is fragile," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 3681–3688.

[11] Laurent Itti, Christof Koch, and Ernst Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[12] Edoardo Ardizzone, Alessandro Bruno, and Giuseppe Mazzola, "Saliency based image cropping," in *Image Analysis and Processing – ICIAP 2013*, 2013, vol. 8156 LNCS, pp. 773–782.

[13] Matthias Kummerer, Thomas S.A. Wallis, Leon A. Gatys, and Matthias Bethge, "Understanding Low- and High-Level Contributions to Fixation Prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, vol. 2017-Octob, pp. 4799–4808.

[14] Lucas Theis, Iryna Korshunova, Alykhan Tejani, and Ferenc Huszár, "Faster gaze prediction with dense networks and fisher pruning," 2018.

[15] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[16] Haya Brama and Tal Grinshpoun, "Heat and Blur: An Effective and Fast Defense Against Adversarial Examples," *arXiv preprint arXiv:2003.07573*, 2020.

[17] Ali Borji and Laurent Itti, "CAT2000: A Large Scale Fixation Dataset for Boosting Saliency Research," *CVPR 2015 workshop on "Future of Datasets"*, 2015.

[18] 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 (ICCV)*, 2017, pp. 618–626.