# Towards Semantic Adversarial Perturbations

**Daniel Saragih**
University of Toronto
daniel.saragih@mail.utoronto.ca

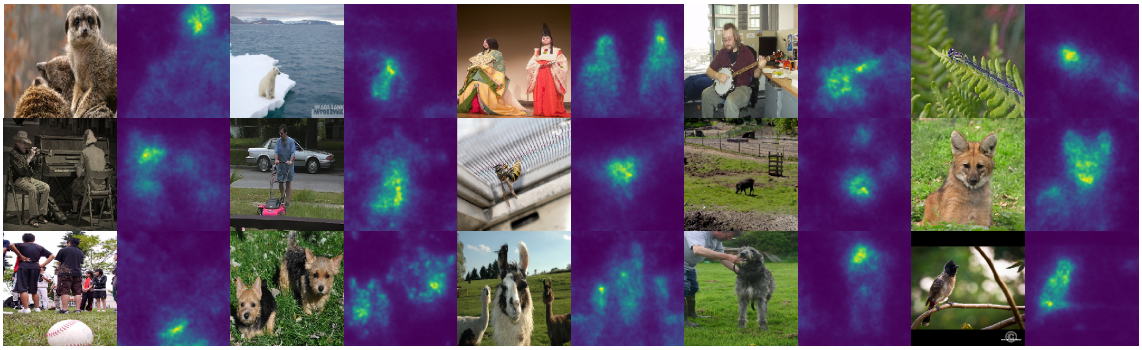

Figure 1: Perceptual Perturbations on typical images taken from ImageNet as explanations. Figure obtained from Elliott et al. [1].

## 1 Summary: Explaining Classifiers Using Adversarial Perturbations on the Perceptual Ball

### 1.1 Methodology

The object of the paper by Elliott et al. [1] is to modify the typical algorithm for adversarial perturbation of classifiers by prefering perturbations on foreground objects and regions of interest (ROI). The method considered is rather simple, suppose we have an image classifier $C(\cdot)$ that takes an image $\mathbf{x}$ and returns a $k$-dimensional vector of probabilities. If the classifier assigns the class $i = \arg\max_j C_j(\mathbf{x})$, then consider

$$M_i(\mathbf{x}') = C_i(\mathbf{x}') - \max_{j \neq i} C_j(\mathbf{x}').$$

Clearly, we'd like to make $M_i(\cdot) \leq 0$. This would yield a sort of untargeted adversarial perturbation.

To restrict the adversarial region to the ROI, the authors simply added a perceptual loss that comes from the ReLU layers of the classifier. Indeed, if we let $C^{(l)}(\cdot)$ be the response at layer $l$, we put together the label loss $M_i(\cdot)$, the prior loss $\|\mathbf{x} - \mathbf{x}'\|_2$, and the perceptual loss to obtain

$$\mathcal{L}(\mathbf{x}') = \lambda_1 M_i(\mathbf{x}') + \lambda_2 \|\mathbf{x} - \mathbf{x}'\|_2 + \lambda_3 \sum_l \left\| C^{(l)}(\mathbf{x}) - C^{(l)}(\mathbf{x}') \right\|_2^2. \quad (1)$$

Adversarial perturbations aim for minimal changes with maximum impact on the loss but often result in off-manifold distortions. Ideally, perturbations would lie on the image manifold to better resemble natural images and reflect the smoothness of the true label distribution. The authors attribute off-manifold issues to *exploding gradients*—exponential growth in layer activations with network depth despite small input changes. To address this, they propose penalizing exploding gradients (i.e., large activation changes) directly, as shown in (1), rather than approximating the image manifold.

### 1.2 Application to Explainability

Qualitative results in Fig. 1 and Fig. 2 demonstrate the method's focus on ROIs, even for challenging images. For instance, the dragonfly on the fern and the baseball in the foreground are isolated from surrounding clutter. This focus enables more interpretable explanations compared to standard adversarial methods. Additionally, extensive ablations on perceptual loss and related comparisons further validate the approach.

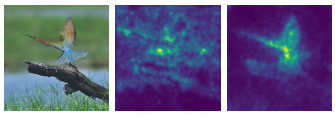

Figure 2: Left to right: Original image; Peturbations without perceptual loss; Magnitude of the perceptual perturbations.

## 2 Experimentation

The semantics of an image refers to the meaning of the object and scene contained therein. Modifying image semantics is in general a challenging problem due to the high dimensionality of the image space and the lack of a simple metric that captures the "understanding" of a scene. To approximate this understanding, we use universal image features trained in a self-supervised manner, specifically DINOv2 [3]. Intuitively, DINO should only register changes on the image manifold while ignoring noise-like perturbations as they are not semantically meaningful. Therefore, semantically meaningful perturbations should cause large changes in the DINO features and we direct these changes with the label loss

$$\mathcal{L}_{label}(\mathbf{x}') = \frac{\lambda_l}{M_\tau(\mathbf{x}) + M_\tau(\mathbf{x}') + 1}$$

where $\tau$ is the target class ID and we use reciprocals to ensure bounds on the loss. Furthermore, we add a term for BRISQUE loss [2] which is a measure of "naturalness" of an image. The final loss is

$$\mathcal{L}_\in(\mathbf{x}') = \mathcal{L}(\mathbf{x}') + \mathcal{L}_{label}(\mathbf{x}') + \lambda_b \text{BRISQUE}(\mathbf{x}') + \frac{\lambda_d}{\|e_{CLS}(x) - e_{CLS}(x')\|_2^2 + 1}, \quad (2)$$

where $e_{CLS}$ are from the DINOv2 CLS tokens. Scalar weights, target ID and hyperparameters can be found in scripts provided in the repository.[1]

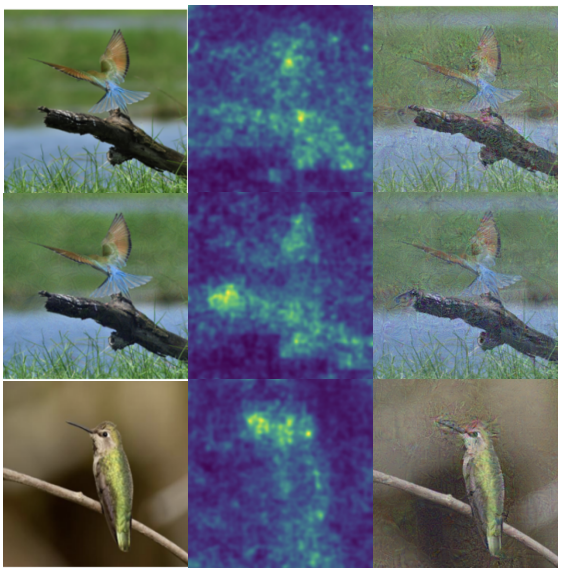

Figure 3: Perturbations using (2). Left to right: ground truth; saliency map; perturbation. Top to bottom: bee killer; adversarial image with PGD; hummingbird.

The results in Fig. 3 show that the perturbations, although much more perceptible than that of the original paper, is still not semantically meaningful. The perturbations are mostly noise-like and unfocused with some new artifacts here and there such as: the red near the bird's feet in the first row and the hooked end of the branch in the second row.

It is therefore natural to ask why the method failed. We list three possible contributors:

1. The base image is a very strong prior. A lot of the "loss budget" is spent to *destroy* the original image and not to *create* a new one.
2. The losses do not provide a strong enough signal towards the target. In particular, $\mathcal{L}_{label}$ seeks the minimal change to the target class.
3. Maximizing distance, as is done in the label and DINO loss, is an ill-posed problem with lots of local maxima.

Due to space constraints, we will explore just the first two points. To test the first point, we start with a base image that is conducive to the target class. In our case, our target class was *goldfish*, so we started with an image of fish in water. We also tried a monochrome image, but the results were not significantly different.

The results in Fig. 4 show that the perturbations still aren't meaningful. If one looks closely, there are orange artifacts on the fish's body and the water which shows the losses working, however, they seem unfocused. Moreover, as we increase the dino scalar $\lambda_d$, the DINO loss seems more focused at changing the brightness of the image rather than the semantics. This perhaps hits at the second and third point: the label loss are not providing a strong enough signal – indeed one can use PGD to generate adversarials from monochrome images – and the DINO features seem to have found a local maxima that is not semantically meaningful.

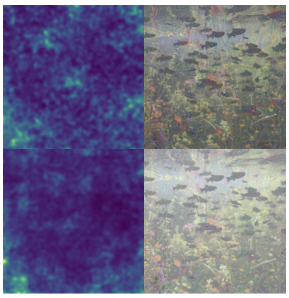

Figure 4: Perturbations on fish images. Top to bottom: increasing $\lambda_d$ values.

Now, we consider the second point. The result in Fig. 5 shows that the DINO loss is focused on destroying the most salient part of the image and all the label loss does is to add adversarial noise, like PGD, on top of it. This indicates that the losses talk past each other leading to suboptimal results.

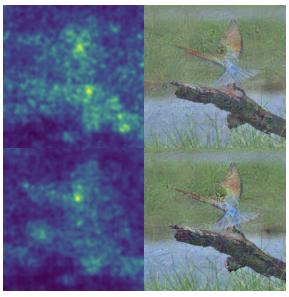

Figure 5: Top image: perturbations using both label and DINO loss. Bottom image: perturbations using just DINO loss.

[1] https://github.com/dsaragih/perceptualBall

## References


[1] Andrew Elliott, Stephen Law, and Chris Russell. Explaining Classifiers using Adversarial Perturbations on the Perceptual Ball. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10688–10697, Nashville, TN, USA, June 2021. IEEE. ISBN 978-1-66544-509-2. doi: 10.1109/CVPR46437.2021.01055.

[2] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, December 2012. ISSN 1941-0042. doi: 10.1109/TIP.2012.2214050.

[3] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, February 2024.