Object Localisation using Perturbations on the Perceptual Ball

Stephen Law [1 2 *] **Andrew Elliott** [1 *] **Chris Russell** [3]

[*]Equal contribution

Abstract

We present a simple regularisation of adversarial perturbations based upon the perceptual loss. While the resulting perturbations remain imperceptible to the human eye, they differ from existing adversarial perturbations in that they are semi-sparse alterations that highlight objects and regions of interest while leaving the background unaltered. We demonstrate the merits of our approach by evaluating on a standard explainability benchmark for object localisation. As a semantically meaningful adverse perturbations, it forms a bridge between counterfactual explanations and adversarial perturbations in the space of images.

We address the gap between counterfactual explanations (Wachter et al., 2017) and adversarial perturbations (Szegedy et al., 2013), and attempt to understand why a minimal changes in image data that results in a change in classifier response does not result in semantically meaningful alteration. One might hope that the smallest edit to change the classifier response of an image labeled as bird should alter the bird pixels, but in practice adversarial perturbations make non-local changes that break the classifier. We show how penalising changes in the mid-level classifier response with a perceptual loss stop this breakage and instead results in semantically meaningful changes that highlight the extent of objects in images (see Fig. 2).

The close relationship between adversarial perturbations and counterfactual explanations follows from the definitions in philosophy and folk psychology of a counterfactual explanation as answering the question "What would need to be different for another outcome to have occurred?" With full causal models of images being outside our grasp, such questions are commonly answered using the Closest Possible World of Lewis (1973), rather than Structured Causal Models of Pearl (2000). Under Lewis's framework, an explanation for why an image is classified as 'dog' can be found by searching for the most similar possible world (i.e. image) which is assigned a different label.

[1]The Alan Turing Institute [2]University College London [3]Amazon. Some work done at 1 and the University of Surrey. Correspondence to: Andrew Elliott <aelliott@turing.ac.uk>, Stephen Law <slaw@turing.ac.uk>, Chris Russell <cmruss@amazon.de>.

2020 ICML Workshop on Human Interpretability in Machine Learning (WHI 2020).

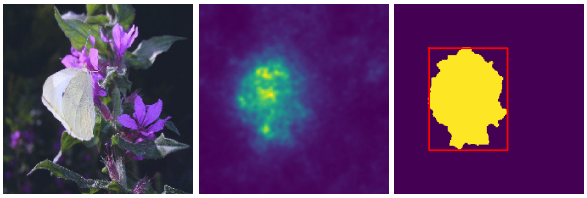

Figure 1. Object localisation. From left to right: Original image; Magnitude of the perceptual perturbations; Dominant connected component and the resulting bounding box from automatic object detection. Our method ignores the flowers and highlights the cabbage butterfly as the most relevant.

One argument for why adversarial perturbations are imperceptible, rather than corresponding to semantically meaningful counterfactual explanations, attributes the effectiveness of adversarial perturbations to exploding gradients. This is the phenomenon where changes in functional response grow exponentially with the depth of the network, relative to a change of input of fixed magnitude. These exploding gradients are an issue known to afflict the learning of Recurrent Neural Networks (Pascanu et al., 2012), and the deep networks common to computer vision. This phenomenon occurs because, by construction, neural networks form a product of (convolutional) matrix operations interlaced with non-linearities; and for directions/locations in which these non-linearities act approximately linearly, the eigenvalues of the Jacobian can grow exponentially with depth (Pascanu et al., 2012). While this is well-studied in the context of training networks the same phenomena occurs when generating adversarial perturbations. Thus, a carefully chosen small perturbation can have an extremely large effect in the response of a deep or recurrent classifier.

To stop adversarial perturbations from exploiting exploding gradients, we propose a simple novel regularisation that bounds the exponential growth of the classifier response by regularising the perceptual distance (Johnson et al., 2016) between the original image and its adversarial perturbation.

1. Prior work

Many approaches to adversarial perturbations have been proposed. These can loosely be divided into white-box ap-

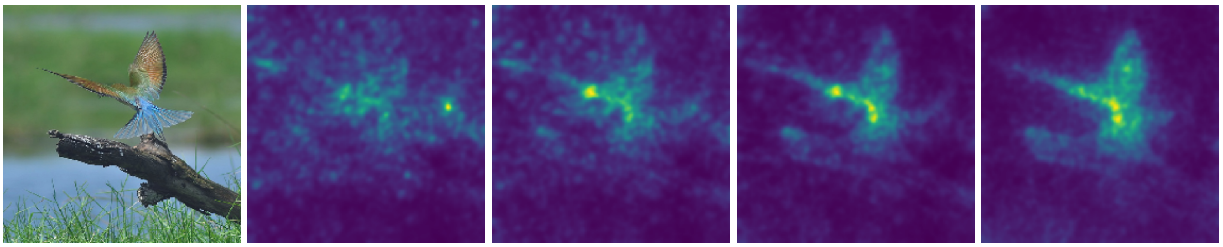

Figure 2. A representative image showing the effects of regularisation over different VGG19_bn layers. As we extend the regularisation to cover higher layers we find the perturbation becomes more compact and better localised upon the object.

proaches e.g. (Carlini & Wagner, 2017; Moosavi-Dezfooli et al., 2016) that assume access to the classification model and black-box methods which do not e.g. (Liu et al., 2017). It is often formulated as trying to find the closest point to an image, under the ℓ_∞, ℓ_1 or ℓ_2 norm, that is predicted a different label.

Adversarial Perturbations and Counterfactuals Conceptually, counterfactual explanations are no different to searching for an adversarial perturbation sampled from the space of possible images. Several approaches have been proposed that either naïvely ignore the requirement that the world is possible (Wachter et al., 2017), use prototypes (Van Looveren & Klaise, 2019), or auto-encoders (Dhurandhar et al., 2018) to characterise the manifold of plausible images, or require large edits that replace regions of the image, either with the output of GANs (Chang et al., 2019) or with patches from other images (Goyal et al., 2019).

Adversarial Perturbations and Gradient Methods The majority of computer vision explainability methods tend to be gradient-based and assign an importance weight to either: every pixel; every super-pixel; or to a set of mid-level neurons. These gradient methods and adversarial perturbations are strongly related. In fact, with most modern networks being piecewise linear, if the found adversarial perturbation and the original image lie on the same linear piece, the difference between the original image and closest adversarial perturbation under the ℓ_2 norm will be a scaled multiple of the gradient. As such, ℓ_2 adversarial perturbations can be thought of as a slightly robustified method of estimating the gradient, that takes into account some local non-linearities.

Of the pure gradient-based approaches, Simonyan et al. (2013) calculated the output gradient with respect to the input image to create a saliency map giving fine-grained, but potentially less interpretable results. Other gradient approaches include SmoothGrad (Smilkov et al., 2017) and Integrated Gradients (Sundararajan et al., 2017). CAM based approaches, such as GradCAM (Selvaraju et al., 2016), combine the gradients and the activation maps in the last convolutional layer of the network to create heatmaps that highlight the salient regions. Perturbation-based methods estimate the local sensitivity over a larger range than gradient methods. Zeiler & Fergus (2014) applied constant-value occlusion masks to different input patches repeatedly to find sensitive regions. Recent work on Extremal Perturbation (Fong & Vedaldi, 2017) estimated a mask to occlude that should have a maximal effect on the network's output.

2. Methodology

We consider a classifier $C(\cdot)$ that takes an image x as input, and returns a k dimensional confidence vector. The classifier $C(\cdot)$ assigns the label $i = \arg\max_j C_j(x)$ to the image x.

Given image x classified as label i we consider the scalar multi-class margin:

$$M_i(x') = C_i(x') - \max_{j \neq i} C_j(x') \tag{1}$$

and note that $M_i(x') \leq 0$ if and only if $C(\cdot)$ does not assign label i to image x. As such an adversarial perturbation x' can be found by minimising:

$$\left(M_i(x') - T\right)^2 \tag{2}$$

where T is a small target value greater than zero. It is well-known (Sorensen, 1982) that minimising a loss of the form:

$$\left(M_i(x') - T\right)^2 + \lambda ||x' - x||_2^2 \tag{3}$$

is equivalent to finding a minimiser of Eq. (2) subject to the requirement that x' lies in the ball defined by $||x - x'||_2^2 \leq \rho$ for some ρ. As such minimising this objective for an appropriate value of λ and T is a good strategy for finding adversarial perturbations of image x with small ℓ_2 norm.

Writing $C^{(l)}(x)$ for the classifier response of in the l^{th} layer of the neural net, we consider the related loss:

$$\left(M_i(x') - T\right)^2 + \lambda' \sum_{l \in \mathcal{L}} ||C^{(l)}(x)' - C^{(l)}(x)||_2^2 + \lambda ||x' - x||_2^2 \tag{4}$$

defined over a set of layers of the neural net $\mathcal{L}$.

The second term of this objective is the perceptual loss of Johnson et al. (2016), and minimising this objective is equivalent to finding a minimiser of Eq. (3) subject to the requirement that x' lies in the ball defined by $\sum_{l \in \mathcal{L}} ||C^{(l)}(x) - C^{(l)}(x)'||_2^2 \leq \rho'$ for some ρ'.

Figure 3. Perceptual Perturbations on ImageNet as explanations. Illustration of perceptual perturbations on typical images taken from ImageNet (Russakovsky et al., 2015). See discussion in Section 3.

Method	Error
GradCAM (Selvaraju et al., 2016)	0.47
guided-GradCAM (Selvaraju et al., 2016)	0.46
SmoothGrad (Smilkov et al., 2017)	0.46
IntegratedGrad (Sundararajan et al., 2017)	0.44
Excitation (Zhang et al., 2017)	0.45
Extremal (Fong et al., 2019)	0.53
GuidedBP (Springenberg et al., 2014)	0.46
RISE (Petsiuk et al., 2018)	0.57
DFool (Moosavi-Dezfooli et al., 2016)	0.57
Us Unguided	*0.43*
Us Guided	**0.41**

Table 1. Object localisation error for the best thresholding strategy for each method. Our methods achieve the lowest error.

3. Perceptual Perturbations as Explanations

We give a qualitative analysis of the perceptual perturbations, as shown in Fig. 3. The found perturbations do a good job of localising on a single object class, even in the presence of highly textured or cluttered images (dragonfly on fern; coral reef). Some error in localisation seems to arise from supporting classes being adjacent to the object - for example, human legs behind the lawnmower are found to be salient.

We evaluate the quality of our perceptual perturbations as explanations by using the weak localisation protocol of Fong & Vedaldi (2017), and test our approach on the first 1000 ImageNet (Russakovsky et al., 2015) validation images.

We take the per-pixel L2 norm of our perturbation as its salience. We then construct a set of bounding boxes for the most dominant region using three simple thresholding strategies based on: thresholding the raw values, thresholding scaled by the image mean, and thresholding a fixed percent of the image. We apply the same strategies, varying thresholds for all methods and report the per method best score. For each threshold, we extract the largest connected component and draw a bounding box around it. The object is assumed to be successfully localised when the Intersection Over Union measure (IOU) between this box and the ground truth is above 0.5. Following GradCAM's guided version (Selvaraju et al., 2016), which makes use of image gradients, we consider a guided variant of our own consisting of an element-wise multiplication between our perturbations and the normalised gradient of the $C_i(x)$ with respect to the image x.

Through a sensitivity study, we identify a sequential set of ReLU layers to regularise over in a VGG19_bn network using the raw value threshold. For the unguided variant of our method, we regularise the ReLU layers from layer 5 to 10 and for the guided variant of our method, we regularise the ReLU layers from layer 2 to 13. For our two perceptual methods, we set $\lambda' = 10000, \lambda = 1$ in Eq. (4) when testing all three strategies. Qualitative evidence confirms that regu-

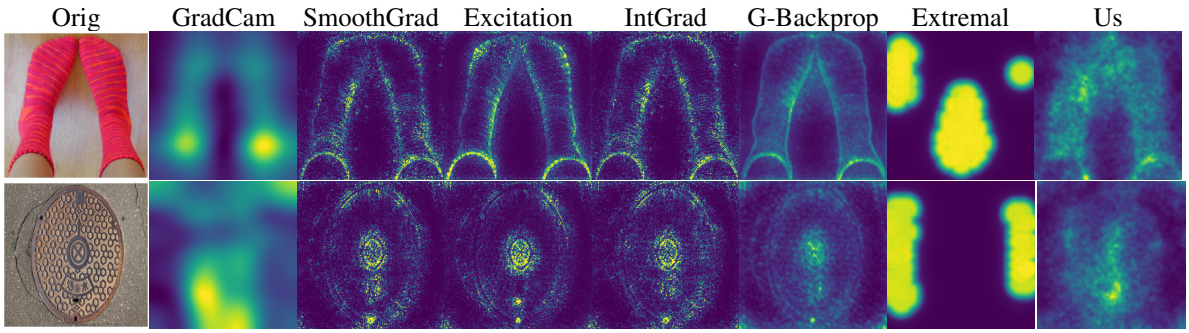

Figure 4. Comparison of explainability methods on ImageNet validation images.

larising sequentially more and higher layers tend to improve object localisation in the image (e.g. see Fig. 2).

We compare our perceptual method and its guided variant with a wide range of alternative approaches (see Table 1 for results and citations). We perform better than all other methods on weak object localisation, where the guided variant and the unguided variant of our method achieve the lowest error and second-lowest error rates respectively. Further, our guided approach also outperforms all others in every thresholding strategy (results not shown).

A qualitative evaluation can be seen in Fig. 4. Our method highlights the interior textures of the target object in the image. This differs from the gradient-based approaches which capture finer edge details and the activation-based GradCam which highlight the entire object coarsely.

4. Conclusion

We have presented a novel regularisation based on the perceptual loss for the generation of adversarial perturbations. This regularisation penalises adversarial perturbations that exploit exploding gradients, forcing larger and more meaningful perturbations to be generated. The fact that such perturbations still exist under these constraints and remain imperceptible to humans is another piece of the puzzle in understanding the interrelationship between adversarial perturbations, neural networks, and human vision. We have shown how these perturbations can be interpreted as explanations and obtained state-of-the-art results on a standard explainability benchmark.

Acknowledgement

This work was supported by the Omidya Group and The Alan Turing Institute under the UK Engineering and Physical Sciences Research Council (EPSRC) grant no. EP/N510129/1 and Accenture Plc. Moreover, we acknowledge Pearl for the computing resources and in particular the help of Tomas Lazauskas and Suleman Tariq.

References

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57. IEEE, 2017.

Chang, C.-H., Creager, E., Goldenberg, A., and Duvenaud, D. Explaining image classifiers by counterfactual generation. In *ICLR*, 2019.

Dhurandhar, A., Chen, P.-Y., Luss, R., Tu, C.-C., Ting, P., Shanmugam, K., and Das, P. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *NeurIPS*, pp. 592–603, 2018.

Fong, R., Patrick, M., and Vedaldi, A. Understanding deep networks via extremal perturbations and smooth masks. *arXiv:1910.08485*, 2019.

Fong, R. C. and Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. *ICCV*, Oct 2017. doi: 10.1109/iccv.2017.371.

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., and Lee, S. Counterfactual visual explanations. *arXiv preprint arXiv:1904.07451*, 2019.

Johnson, J., Alahi, A., and Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pp. 694–711. Springer, 2016.

Lewis, D. *Counterfactuals*. John Wiley & Sons, 1973.

Liu, Y., Chen, X., Liu, C., and Song, D. Delving into transferable adversarial examples and black-box attacks. *ICLR*, 2017.

Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. Deepfool: A simple and accurate method to fool deep neural networks. *CVPR*, Jun 2016. doi: 10.1109/cvpr.2016.282.

Pascanu, R., Mikolov, T., and Bengio, Y. Understanding the exploding gradient problem. *CoRR, abs/1211.5063*, 2, 2012.

Pearl, J. *Causality: models, reasoning and inference*, volume 29. Springer, 2000.

Petsiuk, V., Das, A., and Saenko, K. Rise: Randomized input sampling for explanation of black-box models. *arXiv:1806.07421*, 2018.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *arXiv e-prints*, art. arXiv:1610.02391, Oct 2016.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks. *ICLR*, 2013.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. Smoothgrad: removing noise by adding noise. *arXiv:1706.03825*, 2017.

Sorensen, D. C. Newton's method with a model trust region modification. *SIAM Journal on Numerical Analysis*, 19(2):409–426, 1982.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: The all convolutional net. *arXiv:1412.6806*, 2014.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. *arXiv:1703.01365*, 2017.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv:1312.6199*, 2013.

Van Looveren, A. and Klaise, J. Interpretable counterfactual explanations guided by prototypes. *arXiv:1907.02584*, 2019.

Wachter, S., Mittelstadt, B., and Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the gpdr. *Harv. JL & Tech.*, 31: 841, 2017.

Zeiler, M. and Fergus, R. Visualizing and understanding convolutional networks. *ECCV*, 2014.

Zhang, J., Bargal, S. A., Lin, Z., Brandt, J., Shen, X., and Sclaroff, S. Top-down neural attention by excitation backprop. *Int. J. Comput. Vis.*, 126(10):1084–1102, Dec 2017. ISSN 1573-1405. doi: 10.1007/s11263-017-1059-x.