Componentwise Adversarial Attacks

Lucas Beerens and Desmond J. Higham(✉)

School of Mathematics and The Maxwell Institute for Mathematical Sciences,
University of Edinburgh, Edinburgh EH8 9BT, UK
L.Beerens@sms.ed.ac.uk, d.j.higham@ed.ac.uk

Abstract. We motivate and test a new adversarial attack algorithm that measures input perturbation size in a relative componentwise manner. The algorithm can be implemented by solving a sequence of linearly-constrained linear least-squares problems, for which high quality software is available. In the image classification context, as a special case the algorithm may be applied to artificial neural networks that classify printed or handwritten text—we show that it is possible to generate hard-to-spot perturbations that cause misclassification by perturbing only the "ink" and hence leaving the background intact. Such examples are relevant to application areas in defence, business, law and finance.

Keywords: backward error · misclassification · stability

1 Motivation

It is well known that deep learning image classification tools can be vulnerable to *adversarial attacks*. In particular, a carefully chosen perturbation to an image that is imperceptible to the human eye may cause an unwanted change in the predicted class [7,15]. The fact that automated classification tools may be fooled in this way raises concerns around their deployment in high stakes application areas, including medical imaging, transport, defence and finance [11]. Over the past decade, there has been growing interest in the development of algorithms that construct attacks, and strategies that defend against them [1,6,10,12,13]. Amidst the background of this war of attrition, there has also been "bigger picture" theoretical research into the existence, computability and inevitability of adversarial perturbations [2,5,14,16,17].

In this work, we contribute to the algorithm development side of the adversarial attack literature. We focus on the manner in which perturbation size is measured. Figure 1 illustrates the benefits of our new algorithm. On the left, we show the image of a handwritten digit from the MNIST data set [9]. A trained neural network (accuracy 97%) correctly classified this image as a digit 8. In the middle of Fig. 1 we show a perturbed image produced by the widely used

LB was supported by the MAC-MIGS Centre for Doctoral Training under EPSRC grant EP/S023291/1. DJH was supported by EPSRC grants EP/P020720/1 and EP/V046527/1.

L. Iliadis et al. (Eds.): ICANN 2023, LNCS 14254, pp. 542–545, 2023.
https://doi.org/10.1007/978-3-031-44207-0_45

DeepFool algorithm [12]. This perturbed image is classified as a 2 by the network. On the right in Fig. 1 we show another perturbed image, produced by our new algorithm. This new image is also classified as a 2. The Deepfool algorithm looks for a perturbation of minimal Euclidean norm, treating all pixels equally. In this case, we can see that although the perturbed image is close to the original, there are tell-tale smudges to the white background. Our new algorithm seeks a perturbation that causes a minimal componentwise relative change; and in this context it will not make any change to zero-valued pixels. We argue that the perturbation produced is less noticeable to the human eye, being consistent with a streaky pen, rough paper, or irregular handwriting pressure.

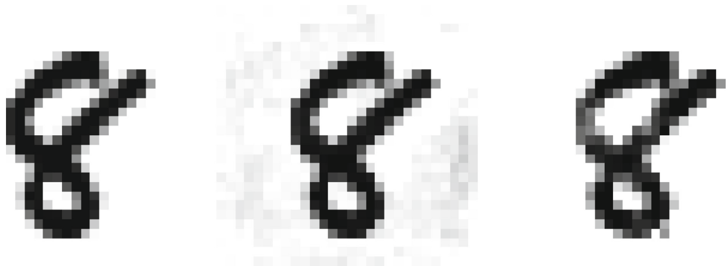

Fig. 1. Showcasing the capabilities of our new algorithm, which seeks a perturbation that causes minimal componentwise relative change. Left: image from the MNIST data set [9], correctly classified as an 8 by a neural network. Middle: perturbed image produced by Deepfool [12], classified as a 2. Right: perturbed image produced by new componentwise algorithm, also classified as a 2. The componentwise algorithm does not change the background, where pixel values are zero. In the notation of Sect. 2, the relative Euclidean norm perturbation size, $\|\Delta x\|_2/\|x\|_2$, is 0.09 for Deepfool and 0.23 for the componentwise algorithm. This reflects the fact that Deepfool looks for the smallest Euclidean norm perturbation whereas the componentwise algorithm has a different objective.

2 Overview of Algorithm

We will focus on image classification, assuming that there are c possible classes. Regarding an image as a normalized vector in $x \in \mathbb{R}^n$, a classifier takes the form of a map $F : [0, 1]^n \to \mathbb{R}^c$, where we assume that output class is determined by the largest component of $F(x)$.

Suppose $F(x) = y$ and we wish to perturb the image to $x + \Delta x$ with $F(x + \Delta x) = \widehat{y}$, where the desired output $\widehat{y}$ produces a different classification, so $\widehat{y}$ has a maximum component in a different position to the maximum component of y. In the *untargeted* case, $\widehat{y}$ may be any such vector. In the *targeted* case, we wish to specify which component of $\widehat{y}$ is maximum.

Because we seek a small perturbation, we will use the linearization $F(x + \Delta x) - F(x) \approx \mathcal{A}\Delta x$, where $\mathcal{A} \in \mathbb{R}^{c \times n}$ is the Jacobian of F at x, and F is assumed to be differentiable in a neighbourhood of x. Then, motivated by the connection to (norm-based) backward error developed in [4] and also by the concept of componentwise backward error introduced in [8], we consider the optimization problem

$$\min\{\epsilon : \mathcal{A}\Delta x = \widehat{y} - y, \quad |\Delta x|_i \leq \epsilon f_i \quad \text{for} \quad 1 \leq i \leq n\}. \tag{1}$$

Here $f \geq 0 \in \mathbb{R}^n$ is a given tolerance vector, and we note that choosing $f_i = |x_i|$ forces zero pixels to remain unperturbed. Following the approach in [8] it is then useful to write $\Delta x = Dv$, where $D = \mathrm{diag}(f)$ and $v \in \mathbb{R}^n$ so that our optimization becomes

$$\min\{\|v\|_\infty : \mathcal{A}Dv = \widehat{y} - y\}. \tag{2}$$

In practice, we found that the problem (2) encourages all components of v to achieve the maximum $\|v\|_\infty$, leading to adversarial perturbations that were quite noticeable. We found more success after replacing (2) by

$$\min\{\|Dv\|_2 : \mathcal{A}Dv = \widehat{y} - y\}. \tag{3}$$

Because $\Delta x = Dv$, in this formulation we retain the masking effect where zero values in the tolerance vector f force the corresponding pixels to remain unperturbed. We found that minimizing $\|Dv\|_2$ rather than $\|v\|_\infty$ produced perturbations that appeared less obvious, and this was the approach used for Fig. 1.

It can be shown that the underlying optimization task arising from this approach may be formulated as a linearly-constrained linear least-squares problem. To derive an effective algorithm, various additional practical steps were introduced; notably, (a) projecting to ensure that perturbations do not send pixels out of range, and (b) regarding each optimization problem as a means to generate a direction in which to take a small step within a more general iterative method.

In our presentation, we will show computational results on a range of data sets that illustrate the performance of the algorithm and compare results with state-of-the-art norm-based attack algorithms. We will also explain how a relevant componentwise condition number for the classification map gives a useful warning about vulnerability to this type of attack.

For full details we refer to [3].

References

1. Akhtar, N., Mian, A.: Threat of adversarial attacks on deep learning in computer vision: a survey. IEEE Access **6**, 14410–14430 (2018). https://doi.org/10.1109/ACCESS.2018.2807385
2. Bastounis, A., Hansen, A.C., Vlaĉić, V.: The mathematics of adversarial attacks in AI-Why deep learning is unstable despite the existence of stable neural networks. arXiv:2109.06098 [cs.LG] (2021)

3. Beerens, L., Higham, D.J.: Adversarial ink: Componentwise backward error attacks on deep learning. IMA J. Appl. Math. (2023). https://doi.org/10.1093/imamat/hxad017
4. Beuzeville, T., Boudier, P., Buttari, A., Gratton, S., Mary, T., Pralet, S.: Adversarial attacks via backward error analysis, December 2021. Working paper or preprint. https://ut3-toulouseinp.hal.science/hal-03296180. https://ut3-toulouseinp.hal.science/hal-03296180v3/file/Adversarial_BE.pdf. hal-03296180. Version 3
5. Fawzi, A., Fawzi, O., Frossard, P.: Analysis of classifiers' robustness to adversarial perturbations. Mach. Learn. **107**, 481–508 (2018)
6. Goodfellow, I.J., McDaniel, P.D., Papernot, N.: Making machine learning robust against adversarial inputs. Commun. ACM **61**(7), 56–66 (2018). https://doi.org/10.1145/3134599
7. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, San Diego, CA (2015). arxiv.org/abs/1412.6572
8. Higham, D.J., Higham, N.J.: Backward error and condition of structured linear systems. SIAM J. Matrix Anal. Appl. **13**(1), 162–175 (1992). https://doi.org/10.1137/0613014
9. LeCun, Y., Cortes, C.: MNIST handwritten digit database (2010). http://yann.lecun.com/exdb/mnist/
10. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: 6th International Conference on Learning Representations, Vancouver, BC. OpenReview.net (2018). http://openreview.net/forum?id=rJzIBfZAb
11. Marcus, G.: Deep learning: A critical appraisal. arXiv:1801.00631 [cs.AI] (2018)
12. Moosavi-Dezfooli, S., Fawzi, A., Frossard, P.: DeepFool: a simple and accurate method to fool deep neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, NV, USA, pp. 2574–2582. IEEE Computer Society (2016). https://doi.org/10.1109/CVPR.2016.282
13. Papernot, N., McDaniel, P.D., Goodfellow, I.J., Jha, S., Celik, Z.B., Swami, A.: Practical black-box attacks against machine learning. In: Karri, R., Sinanoglu, O., Sadeghi, A., Yi, X. (eds.) Proceedings of the ACM Conference on Computer and Communications Security, Abu Dhabi, UAE, pp. 506–519. ACM (2017). https://doi.org/10.1145/3052973.3053009
14. Shafahi, A., Huang, W., Studer, C., Feizi, S., Goldstein, T.: Are adversarial examples inevitable? In: International Conference on Learning Representations, New Orleans, USA (2019)
15. Szegedy, C., et al.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
16. Tyukin, I.Y., Higham, D.J., Gorban, A.N.: On adversarial examples and stealth attacks in artificial intelligence systems. In: 2020 International Joint Conference on Neural Networks, pp. 1–6. IEEE (2020)
17. Tyukin, I.Y., Higham, D.J., Bastounis, A., Woldegeorgis, E., Gorban, A.N.: The feasibility and inevitability of stealth attacks. arXiv:2106.13997 (2021)