# With friends like these, who needs adversaries?

Saumya Jetley*, Nicholas A. Lord* and Philip H.S. Torr

{sjetley, nicklord, phst}@robots.ox.ac.uk

## Synopsis

We analyse the responses of deep convolutional image classification networks to systematic changes to their input images, learning not only about the nature of their adversarial vulnerability, but also something more fundamental about their classification behaviour.

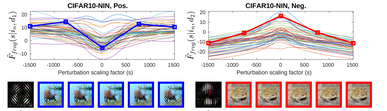



Figure 1: **Adding or subtracting components along $\mathbf{d}_i$ (a specific direction of perturbation in the input image space) causes the network to change its prediction to 'frog'; observe that a 'deer' with a mild diamond striping added to it gets classified as a 'frog'. This trend is fairly consistent across different choices of $\mathbf{i}$, itself, as can be visualised via the individual curves in the plots above where each curve is associated with a randomly sampled $\mathbf{i}_c$. Likewise, perturbations along $\mathbf{d}_i$ change *any* 'frog' to a 'non-frog' class: observe the predicted labels for the sample images along the red curve in the second plot.** Note that $\mathcal{F}_{frog}(\mathbf{s}|\mathbf{i}_c, \mathbf{d}_i, \theta)$ denotes the output of the layer before softmax for 'frog' class and $\mathbf{s}$ is the perturbation scaling factor.

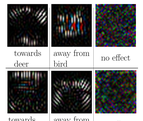



Figure 2: The framework developed in this paper can identify a family of **patterns that move predictions towards or away from any given class, or have no effect on class identity at all**, for any given network.

Key Results:

- These **networks associate specific input image-space directions with fixed class identities**, with little regard for context. This provides a novel perspective on Universal Adversarial Perturbations [6], specifically why they empirically target particular classes.
- The **adversarial vulnerability of DCNs is closely entwined with their performance capabilities**: the input image-space directions along which the networks are most vulnerable to attack are the same directions which they use to achieve their classification performance in the first place.
- **Naive compression-based defence strategies remain vulnerable** to appropriately designed attacks. The vulnerability is fundamental and can only be remedied through development of a net with a substantially different concept of class identity than exists presently.

## Analysis

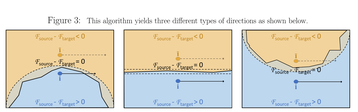
We base our work on the geometric decision boundary analysis of [8], and begin by extracting the mean principal directions and principal curvatures of the classifier's image-space class decision boundaries over the dataset, as outlined below.

**Algorithm 1** Computes mean principal directions and principal curvatures for a net's image-space decision surface.

**Input:** network class score function $\mathcal{F}$, dataset $\mathbf{I} = \{\mathbf{i}_1, \mathbf{i}_2, \cdots, \mathbf{i}_N\}$, target class label $c$
**Output:** principal curvature basis matrix $\mathbf{V_S}$ and corresponding principal curvature vector $\mathbf{v_s}$
**procedure** PrincipalCurvatures($\mathcal{F}$, $\mathbf{I}$, $c$)
  $\mathbf{H} \leftarrow$ null
  **for** each sample $\mathbf{i}_k \in \mathbb{I}$ s.t. $\text{argmax}_l(\mathcal{F}_l(\mathbf{i}_k)) \neq c$ **do**
    $\hat{c} \leftarrow \text{argmax}_l(\mathcal{F}_l(\mathbf{i}_k))$ ▷ network predicts $\mathbf{i}_k$ to be of class $\hat{c}$
    $\mathcal{H}_{\hat{c}c}$: define as Hessian of function $(\mathcal{F}_{\hat{c}} - \mathcal{F}_c)$ [1] ▷ subscripts select class scores
    $\hat{\mathbf{i}}_k \leftarrow$ DeepFool($\mathbf{i}_k$, $c$) ▷ approximate nearest boundary point to $\mathbf{i}_k$
    $\mathbf{H} \leftarrow \mathbf{H} + \mathcal{H}_{\hat{c}c}(\hat{\mathbf{i}}_k)$ ▷ accumulate Hessian at sample boundary point
  $\mathbf{H} \leftarrow \mathbf{H}/\|\mathbb{I}\|$ ▷ normalise mean Hessian by number of samples
  $(\mathbf{V_S}, \mathbf{v_s}) = \text{Eigs}(\mathbf{H})$ ▷ compute eigenvectors and eigenvalues of mean Hessian
  **return** $(\mathbf{V_S}, \mathbf{v_s})$

[1] Note that the classes $\hat{c}$ and $c$ in the algorithm above correspond to the source and target classes of Fig.3 respectively.

Figure 3: This algorithm yields three different types of directions as shown below.




(a) Positive-curvature direction (b) Flat direction (c) Negative-curvature direction

## Existing Hypothesis

- The authors of [8] advance a hypothesis connecting positively curved directions with the universal adversarial perturbations of [6] (see Fig.3(a)).
- They demonstrate that if the normal section of a net's decision surface along a given direction can be locally bounded on the outside by a circular arc of a particular positive curvature in the vicinity of a sample image point, then geometry accordingly dictates an upper bound on the distance between that point and the boundary in that direction.
- If such directions and bounds turn out to be largely common across sample image points (which they do), then the existence of universal adversaries follows directly, with higher curvature implying lower-norm adversaries.

  **It is from this point that we move beyond the prior art and begin an iterative loop of further experimentation and analysis as follows.**

## Our Extended Hypothesis, Associated Experiments and Results

### I - Class identity as a function of the component in specific image-space directions

**Provided that the $2^{nd}$-order boundary approximation holds well over a sufficiently wide perturbation range and variety of images, the model implies that:**

- **The distance of such adversaries from the decision boundary should increase as a function of their norm.**
- **The attack along any *positively* curved direction should increasingly perturb the sample *towards* the corresponding target class: class $c$ in Alg.1.**

Our experimental observations confirm the above conjectures:

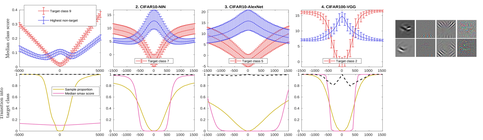



Figure 4: Observe the selected class scores plotted as functions of the perturbation scaling factor $s$ along the most positively curved direction per net. The 'Median class score' plot compares the score of a randomly selected target class with the supremum of the scores for the non-target classes. Each curve represents the median of the class scores over the associated dataset and is bracketed below by the 30th-percentile score and above by the 70th. **Notice that as the perturbation magnitude increases (with either sign), the population's target class score approaches and then surpasses the highest non-target class score.** The 'Transition into target class' plot depicts the fraction of the dataset not originally of the target class, but which is transitioned into the target class by the perturbation. Alongside, we graph that population's median softmax target-class score. The black dashed line represents the fraction of the population originally of the target class that remains in the target class under the perturbation. **Once again, notice the monotonicity in the fraction of non-target samples perturbed into the target class, and in the median target class softmax score, as a function of the perturbation magnitude** $|s|$. The image grid on the right illustrates the 2D visualisations of the two most positively curved directions for randomly selected target classes: the columns correspond, from left to right, with the four net-dataset pairs under study.

**While positively curved directions may be of primary interest in [8], two other important corollaries emerge from an extension of the above geometric argument:**

- **The steps along *negative*-curvature directions perturb increasingly *away* from class $c$.**
- **The plethora of approximately *zero*-curvature (flat) directions identified in [2, 8] should have *negligible* effect on the class identity.**

Once again, the experimental results confirm our hypothesis:

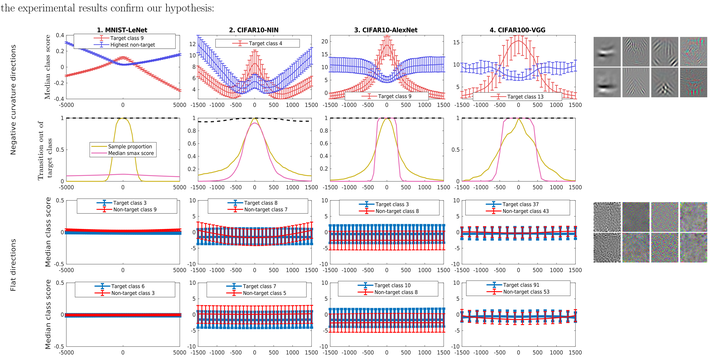



Figure 5: Observe the selected class scores plotted as functions of the scaling factor $s$ of the perturbations along the most negatively curved directions and flat directions per net. The 'Median class score' plot compares the score of a randomly selected target class with the supremum of the scores for the non-target classes, for the negatively curved directions. For the flat curvature directions, it plots the score of a randomly selected target class and non-target class respectively. Each curve represents the median of the class scores over the associated dataset and is bracketed below by the 30th-percentile score and above by the 70th. For the negative-curvature directions, the 'Transition out of target class' graph works in reverse to the corresponding positive-curvature graph in Fig.3 'sample proportion' represents the fraction of the dataset originally of the target class which retains the target-class label under perturbation, with the median softmax target-class score as before. The black dashed line now represents the fraction of the dataset *not* originally of the target class which remains *outside* of the target class under perturbation. **Once again, notice that with an increasing perturbation norm the population's non-target class scores overtake its target class score, with the natural samples of the target class accordingly being perturbed out of it. Further, the flatness of the decision boundary manifests as flatness of both target and non-target class scores: over a wide range of magnitudes, and these directions do not influence the network in any way.** The images in the rightmost column illustrate a sample of these directions as visual patterns. Each block of eight images corresponds to the label (negative, or flat) to its left, and the two-image columns in each block correspond from left to right with the main four net-dataset pairs under study.

### II - Network classification performance versus effective data dimensionality

**A more intuitive picture of what the networks are actually doing begins to emerge:**

- **The nets are identifying the high-curvature image-space directions as features associated with respective class identities.**
- **The mean decision boundary curvature along such a direction can then be thought of as representing the empirical average width of the feature response window within which a class will be classified as the "inside class", rather than the "outside class" of the curving boundary.**
- **Thus, these directions are what the net relies on *generally* in predicting the classes of images, with the curvatures-cum-sensitivities representing their relative weightings.**
- **Accordingly, it should be possible to disregard the "flat" directions of near-zero curvature without any noticeable change in the nets' predictions.**

The results below confirm our intuition.

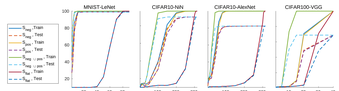



Figure 6: Training and test accuracies of a sample of nets as a function of the subspace onto which their input images are projected. The input subspace is parametrised by a dimensionality parameter $d$, which controls the number of basis vectors selected per class ($d$ varies from 0 until the input space is fully spanned). We use four variants of selection: the $d$ most positively curved directions per class (yielding the subspace $S_{pos}$); the $d$ most negatively curved directions per class (yielding the subspace $S_{neg}$); the union of the previous two (subspace $S_{pos \cup neg}$), and the $d$ least curved (flattest) directions per class (subspace $S_{flat}$). The subspace $\mathcal{S}$ so obtained is represented by the orthonormalised basis matrix $\mathbf{Q}_\mathcal{S}$ (obtained by QR decomposition of the aggregated directions), and each input image $\mathbf{i}$ is then projected onto $\mathcal{S}$ as $\hat{\mathbf{i}}' = \mathbf{Q}_\mathcal{S}\mathbf{Q}_\mathcal{S}^T\hat{\mathbf{i}}$. Note: The mean training-set orthogonal component $(\mathbf{I} - \mathbf{Q}_\mathcal{S}\mathbf{Q}_\mathcal{S}^T)\hat{\mathbf{i}}$ can be added, but is approximately 0 in practice for data normalised by mean subtraction, as is the case here. Observe the relationship between the ordering of curvature magnitudes and classification accuracy by comparing the $S_{flat}$ curves to the others. **The outcome is striking: it is evident that in many cases, classification decisions have effectively already been made based on a relatively small number of features, corresponding to the most curved directions.**

### III - Link between classification and adversarial directions

**Another important point emerges here:**

- **Since it is the high-curvature directions that are largely responsible for determining the nets' classification decisions, the nets should be vulnerable to adversarial attack along *precisely these directions.***

It was noted in [2] that adversarial attack vectors evince high components in subspaces spanned by high-curvature directions. We expand the analysis for various attack methods and confirm the direct relationship between the fraction of adversarial norm in given subspaces and the corresponding usefulness of those subspaces for classification.

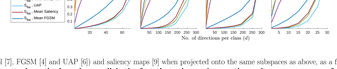



Figure 7: Mean $\ell_2$-norms of various adversarial perturbations (DeepFool [7], FGSM [4] and UAP [6]) and saliency maps [9] when projected onto the same subspaces as above, as a fraction of their original norms. Compare these plots with those in the figure above. **The inclusion of the saliency images of [9] alongside the attack methods makes explicit the fact that adversaries are themselves an exposure of the net's notion of saliency.**

**By now, two results hint at a simpler way of identifying bases of classification/adversarial directions:**

- **The class-score curves sampled and displayed in Fig.1 reveal a direct connection between the curvature of a direction near the origin and its derivative magnitude.**
- **The directions obtained by boundary curvature analysis in Alg.1 correspond to the directions exploited by various *first*-order methods, as in Fig.7.**

Thus, we use a collection of DeepFool perturbations to provide the required gradient information, perform SVD on them, and order the singular vectors by their singular values. As hypothesised, the results neatly replicate the previously seen classification accuracy trends for high-to-low and low-to-high traversal of the singular vectors.

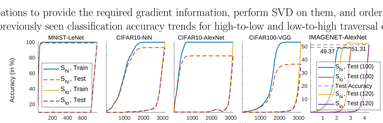



Figure 8: Classification accuracies on image sets projected onto subspaces of the spans of their corresponding DeepFool perturbations. For each net-dataset pair, DeepFool perturbations are computed over the image set and assembled into a matrix that is decomposed into its SVD. The singular vectors are ordered as per their singular values. $S_{hi}$ represents the high-to-low ordering, $S_{lo}$ the low-to-high, and $d$ the number of vectors retained. **Compare this figure to Fig.6 (while noticing how $d$ now counts the total number of directions).** For the ImageNet experiments, owing to memory constraints, the SVD is performed on downsampled DeepFools of size $100 \times 100 \times 3$ and $120 \times 120 \times 3$, respectively. The resulting singular vectors span the entire effective classification space of correspondingly downsampled images. This is evinced by the fact that the classification accuracy of images projected onto the singular vectors' subspace saturates to the same performance as that yielded when the net is tested directly on the downsampled images.

### IV - On image compression and robustness to adversarial attack

**Given the evidence that the effective directions of adversarial attack are also the directions that contribute the most to the DCNs' classification performance, we make the following conjectures:**

- **Any attempt to mitigate adversarial vulnerability by discarding these directions, either by compression of the input data [5, 1, 10] or by suppression of intermediate network representations [3], must effect a loss in the classification accuracy.**
- **Nets must remain just as vulnerable to attack along the remaining classification directions, to the extent that the corresponding class-score functions which possess the properties discussed earlier remain unchanged.**

This is indeed the case, as demonstrated by the results below:

| $d_{in}$ | $E_\infty(\ell_2\ norm(\hat{\mathbf{i}}_k))$ | $E_2(\ell_2\ norm(\hat{\mathbf{i}}_k))$ | Accuracy (%) | Fooling rate (%) | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | $f = 1$ | $f = 2$ | $f = 3$ | $f = 4$ | $f = 5$ |
| 227 | 26798.72 | | 63.96 | 52.75 | 100.00 | 100.00 | 100.00 | 100.00 |
| 200 | 26515.20 | | 53.19 | 55.80 | 32.75 | 77.25 | 88.95 | 92.20 |
| 150 | 26327.03 | | 46.86 | 53.50 | 35.55 | 58.35 | 77.90 | 85.95 |
| 120 | 26139.98 | | 41.92 | 51.75 | 36.15 | 49.80 | 66.30 | 76.90 |
| 100 | 26008.02 | | 37.98 | 48.10 | 41.65 | 49.25 | 59.95 | 68.05 |

Table 1: The images $\hat{\mathbf{i}}_k$ used to train AlexNet operate at the scale of $d_{orig} = 227$ (pixels on a side). In the pre-processing step, these images are downsized to $d_{low}$, before being upsampled back to the original scale. The reconstructed DeepFool perturbations $\hat{\boldsymbol{\delta}}_k$ lose some of their effectiveness, as seen in the fooling-rate column for $f = 1$. **When the effect of downsampling is countered by increasing the value of the $\ell_f$-norms of these perturbations (using higher values of $f$), their efficacy is steadily restored.** Note that the mean norms of images and perturbations are estimated in the upscaled space, as are the classification accuracies. The accuracy values for $d_{low} = \{100, 120\}$ should be compared to those at convergence in Fig.8. Any difference in the performance scores is strictly due to the random selection of the subset of 2000 test images used for evaluation.

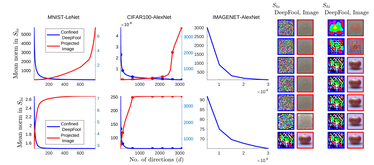



Figure 9: Blue curves depict the mean $\ell_2$-norms of "confined DeepFool" perturbations: those that are calculated under strict confinement to the respective subspaces of Fig.8. Note the differences in scale of the y-axes of the different plots. **Clearly, lower-norm DeepFools can be obtained by restricting the attack's iterative linear optimisation procedure to the space spanned by the "compressed perturbations".** For MNIST and CIFAR, we also plot (in red) the mean norms of the projections of the input images onto those subspaces: observe the inverse relationship between the two curves. The columns on the right visualise, from top to bottom, sample images at the indicated points on the curves in the CIFAR100-AlexNet plots, from left to right: blue-bordered images represent confined DeepFool perturbations (rescaled for display), with their red-bordered counterparts displaying the projection of the corresponding sample CIFAR image onto the same subspace. **Observe that when the human-recognisable object appearance is captured in any given subspace, the corresponding DeepFool perturbation becomes maximally effective (*i.e.* small-norm). Likewise, when the projected image is not readily recognisable to a human, the DeepFool perturbation is large. The feature space *per se* does not account for adversariality: the issue is in the net's response to the features.**

## Conclusion

- We expose a collection of directions along which the net's class-score output functions are nonlinear, but are *de facto* of a relatively constrained form: axis-symmetric and typically monotonic over large ranges, and strikingly similar across the different image samples.
- The way in which DCNs use these features to classify renders them structurally vulnerable to adversarial attack, as it implicitly differs from the way humans solve the same problems.
- For any scheme to be truly effective against the problem of adversarial vulnerability, it must lead to a fundamentally more insightful (and likely complicated) use of features than presently occurs. **Until then, we hope that it is appreciated that as it stands, DCNs' favourite features are their own worst adversaries.**

*Acknowledgements: This work was supported by the EPSRC, ERC grant ERC-2012-AdG 321162-HELIOS, EPSRC grant Seebibyte EP/M013774/1 and EPSRC/MURI grant EP/N019474/1.*

## References

[1] Das, N., Shanbhogue, M., Chen, S., Hohman, F., Chen, L., Kounavis, M.E., Chau, D.H.: Keeping the bad guys out: Protecting and vaccinating deep learning with JPEG compression. CoRR abs/1705.02900 (2017)

[2] Fawzi*, A., Moosavi-Dezfooli*, S.M., Frossard, P., Soatto, S.: Classification regions of deep neural networks. arXiv preprint arXiv:1705.09552 (2017)

[3] Gao, J., Wang, B., Lin, Z., Xu, W., Qi, Y.: Deepcloak: Masking deep neural network models for robustness against adversarial samples. In: International Conference on Learning Representations (2017)

[4] Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: International Conference on Learning Representations (2015). http://arxiv.org/abs/1412.6572

[5] Maharaj, A.V.: Improving the adversarial robustness of convnets by reduction of input dimensionality (2015)

[6] Moosavi-Dezfooli*, S.M., Fawzi*, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. pp. 86–94. IEEE (2017)

[7] Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). No. EPFL-CONF-218057 (2016)

[8] Moosavi-Dezfooli*, S.M., Fawzi*, A., Fawzi, O., Frossard, P., Soatto, S.: Robustness of classifiers to universal perturbations: A geometric perspective. In: International Conference on Learning Representations (2018)

[9] Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)

[10] Xie, C., Wang, J., Zhang, Z., Ren, Z., Yuille, A.L.: Mitigating adversarial effects through randomization. CoRR abs/1711.01991 (2017)