Universal adversarial perturbations

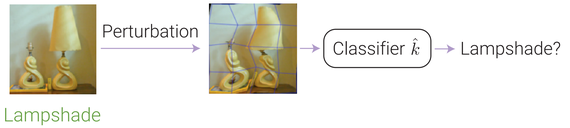
Seyed-Mohsen Moosavi-Dezfooli*, Alhussein Fawzi+, Omar Fawzi†, Pascal Frossard*
*EPFL, +UCLA,†ENS-Lyon

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

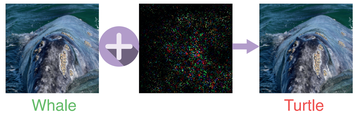
LTS4

Robustness of deep networks

For image classifiers, achieving robustness to perturbations is a crucial requirement

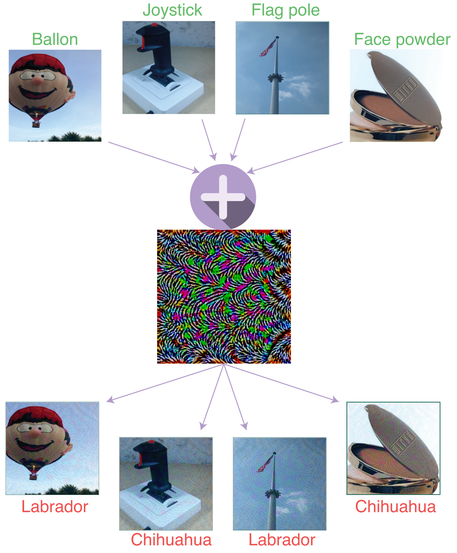


E.g. adversarial perturbations causing misclassification



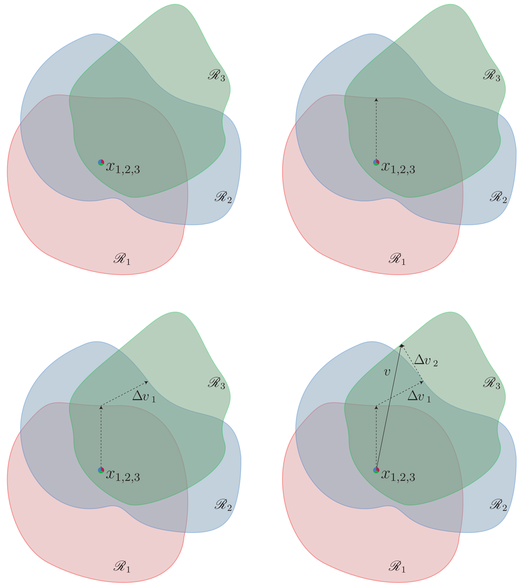
Universal adversarial perturbation

A **single perturbation** causing misclassification with high probability



The algorithm

Finding universal perturbation using the set X



Algorithm Computation of universal perturbations.

input: Data points X, classifier $\hat{k}$, desired ℓ_p norm of the perturbation ξ, desired accuracy on perturbed samples δ.
output: Universal perturbation vector v.
Initialize $v \leftarrow 0$.
while $\text{Err}(X_v) \leq 1 - \delta$ **do**
  for each datapoint $x_i \in X$ **do**
    if $\hat{k}(x_i + v) = \hat{k}(x_i)$ **then**
      Compute the **minimal** perturbation that sends $x_i + v$ to the decision boundary:

$$\Delta v_i \leftarrow \arg\min_r \|r\|_2 \text{ s.t. } \hat{k}(x_i + v + r) \neq \hat{k}(x_i).$$

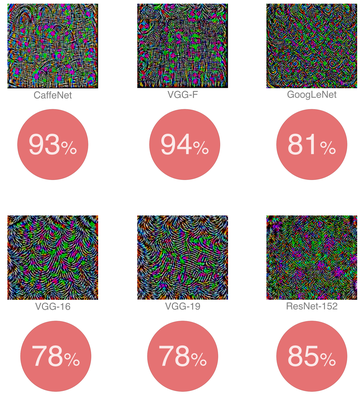
      Update the perturbation:

$$v \leftarrow \mathcal{P}_{p,\xi}(v + \Delta v_i).$$

    end if
  end for
end while

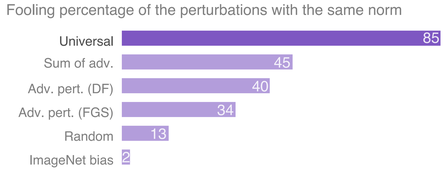
Misclassification rate

High misclassification rate on *unseen* test data



What is special in these perturbations?

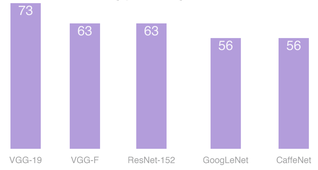
Comparison of universal perturbation to other types of perturbations



Random noise achieving 90% fooling rate

Doubly universal perturbations

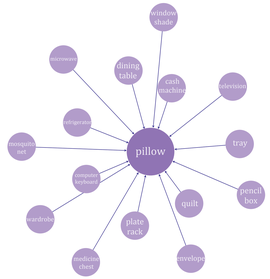
The ability to generalize well across different neural networks



Perturbed labels

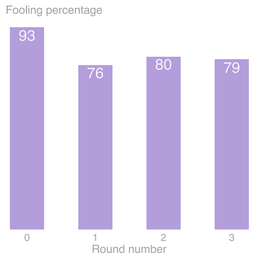
Existence of *dominant* labels that suck most other labels.

High volume classification regions!

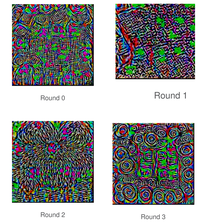


Feedbacking does not help!

Augmenting the training data with perturbed images

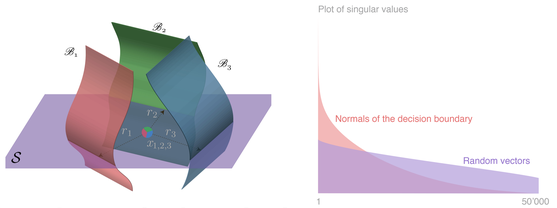


... because universal perturbations are diverse!



Explaining universal perturbations

The decision boundary in the vicinity of natural images is highly correlated!



New theoretical analysis:

Check our new work: “Analysis of universal adversarial perturbations”

“Universal adversarial perturbations”, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, Honolulu, US.

github.com/lts4/universal