Adversarial Examples for Semantic Segmentation and Object Detection

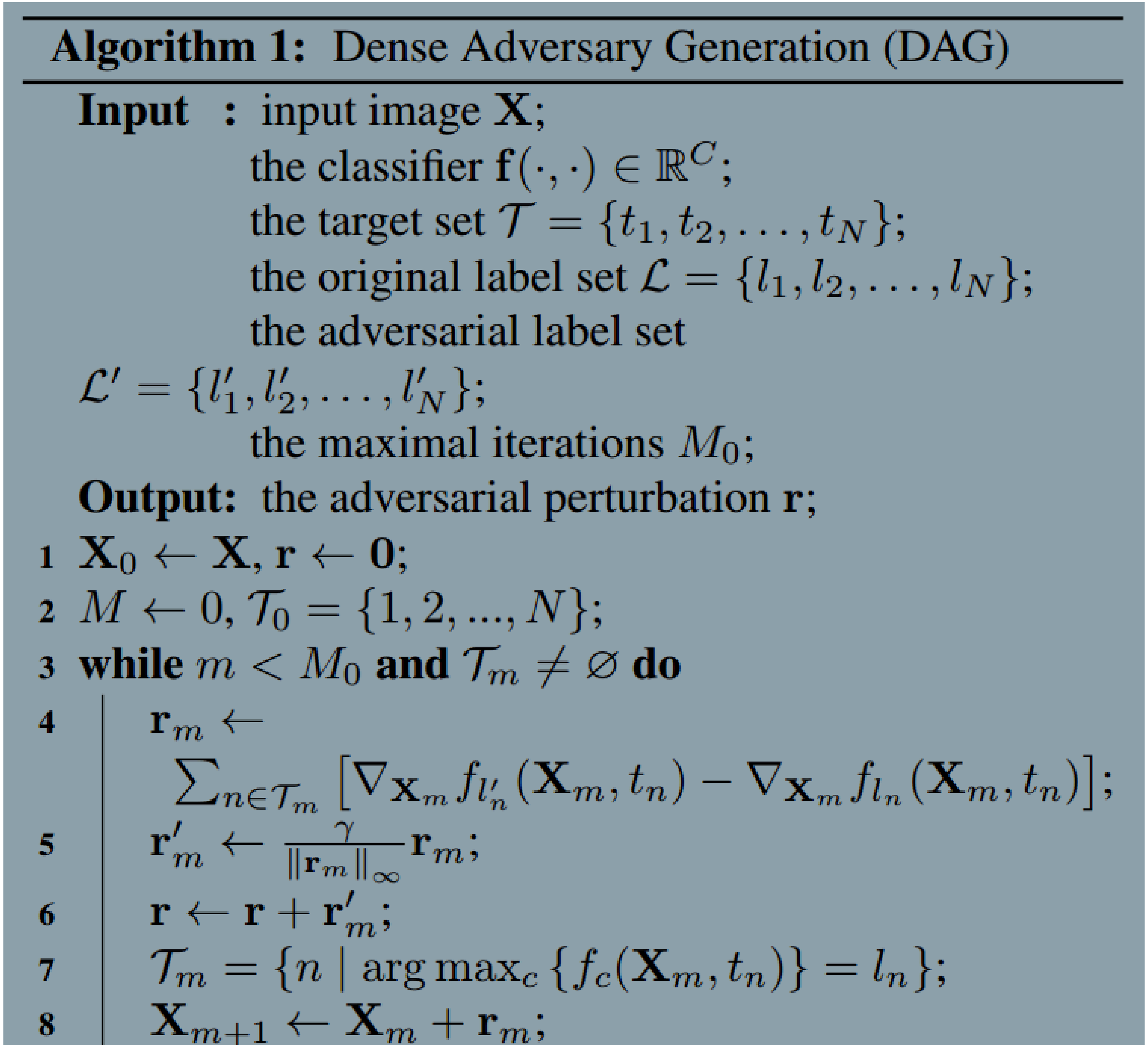
published: 2017-04

authors: Hohns Hopkins University

> It has been well demonstrated that adversarial examples, i.e., natural images with visually imperceptible pertubations added, generally exist for deep networks to fail on image classification. We extend adversarial examples to semantic segmentation and object detection which are much more difficult.

> **Deep networks are often sensitive to small perturbations to the input image. It has shown that adding visually imperceptible perturbations can result in failures for image classification. These perturbed images, often called adversarial examples, are considered to fall on some areas in the large, high-dimensional feature space which are not explored in the training process.**

> Given an image and the recognition targets, DAG generates an adversarial perturbation which is aimed at confusing as many targets as possible.



Algorithm 1: Dense Adversary Generation (DAG)

```
Input  : input image $\mathbf{X}$;
         the classifier $\mathbf{f}(\cdot,\cdot) \in \mathbb{R}^C$;
         the target set $\mathcal{T} = \{t_1, t_2, \ldots, t_N\}$;
         the original label set $\mathcal{L} = \{l_1, l_2, \ldots, l_N\}$;
         the adversarial label set $\mathcal{L}' = \{l'_1, l'_2, \ldots, l'_N\}$;
         the maximal iterations $M_0$;
Output: the adversarial perturbation $\mathbf{r}$;
$\mathbf{X}_0 \leftarrow \mathbf{X}$, $\mathbf{r} \leftarrow \mathbf{0}$;
$M \leftarrow 0$, $\mathcal{T}_0 = \{1, 2, ..., N\}$;
while $m < M_0$ and $\mathcal{T}_m \neq \varnothing$ do
    $\mathbf{r}_m \leftarrow \sum_{n \in \mathcal{T}_m} \left[\nabla_{\mathbf{X}_m} f_{l'_n}(\mathbf{X}_m, t_n) - \nabla_{\mathbf{X}_m} f_{l_n}(\mathbf{X}_m, t_n)\right]$;
    $\mathbf{r}'_m \leftarrow \frac{\gamma}{\|\mathbf{r}_m\|_\infty} \mathbf{r}_m$;
    $\mathbf{r} \leftarrow \mathbf{r} + \mathbf{r}'_m$;
    $\mathcal{T}_m = \{n \mid \arg\max_c \{f_c(\mathbf{X}_m, t_n)\} = l_n\}$;
    $\mathbf{X}_{m+1} \leftarrow \mathbf{X}_m + \mathbf{r}_m$;
```