به نام خدا

Adversarial attack: اگر مقداری مانند r به تصویر اضافه شود به طوری که تصویر جدید به صورت چشمی شبیه به تصویر اصلی باشد ولی کلسیفایر ما را به اشتباه بیندازد و برچسب داده بعد از اضافه کردن این مقدار r (که به آن perturbation گفته میشود) متفاوت از برچسب بدون r شود. مانند شکل زیر:

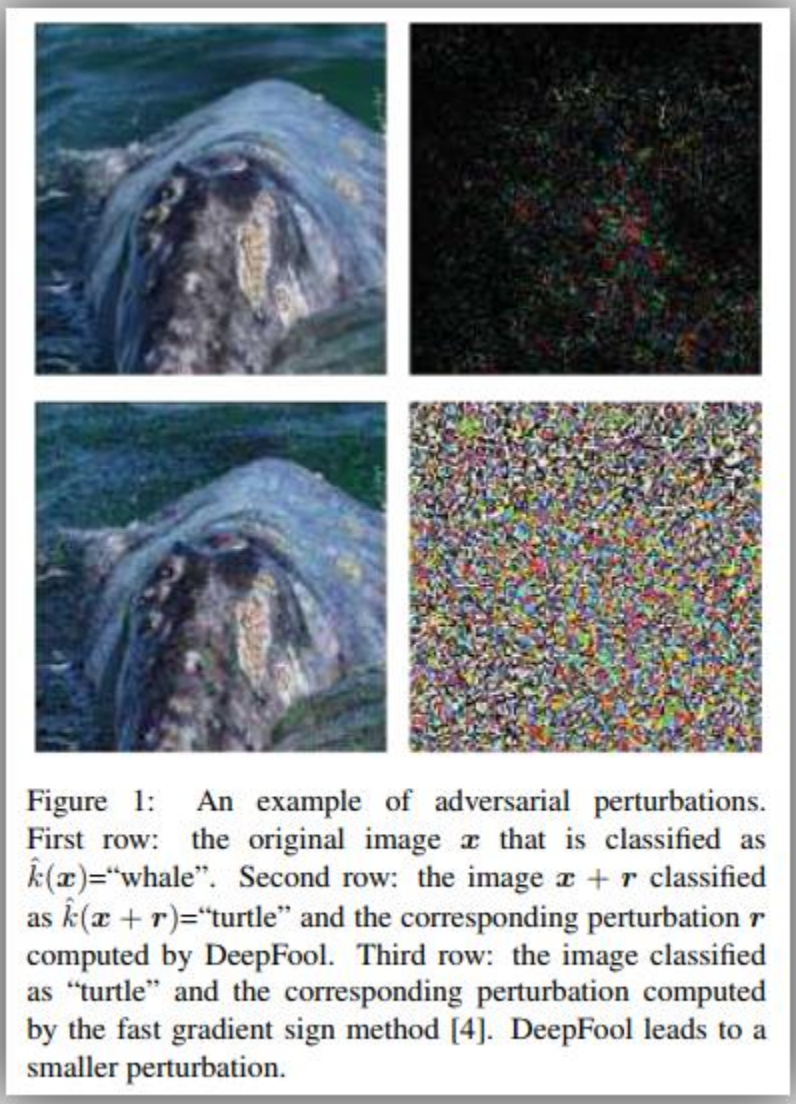

Figure 1: An example of adversarial perturbations. First row: the original image $\boldsymbol{x}$ that is classified as $\hat{k}(\boldsymbol{x})$="whale". Second row: the image $\boldsymbol{x} + \boldsymbol{r}$ classified as $\hat{k}(\boldsymbol{x} + \boldsymbol{r})$="turtle" and the corresponding perturbation $\boldsymbol{r}$ computed by DeepFool. Third row: the image classified as "turtle" and the corresponding perturbation computed by the fast gradient sign method [4]. DeepFool leads to a smaller perturbation.

برای مقابله با این مشکل روشی به اسم DeepFool ارائه شده است که به مختصر درباره آن توضیح میدهم:

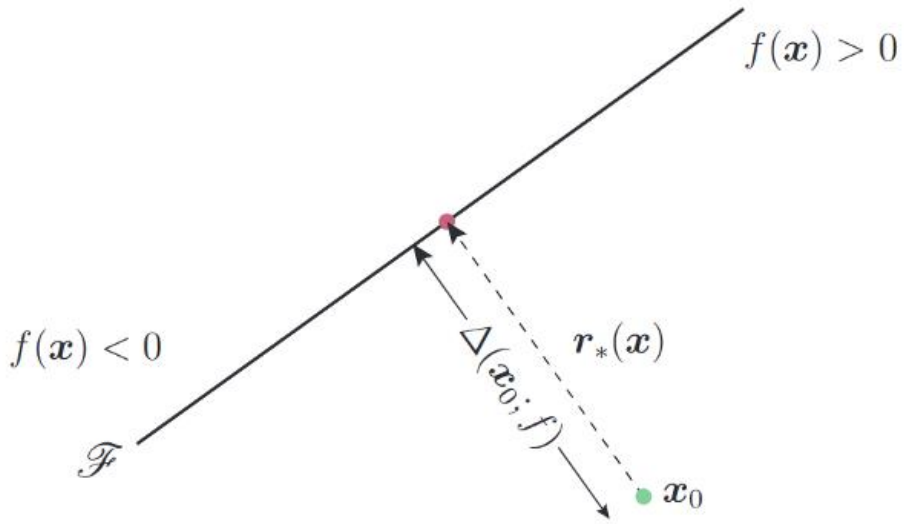

Figure 2: Adversarial examples for a linear binary classifier.

Robustness این مدل فاصله x0 از خطی است که دو کلاس را از هم جدا میکند.

کوچکترین perturbation ای که تصمیم کلسیفایر را عوض میکند برابر با پروجکشن اورتوگونال(ساخت زاویه عمود با خط کلسیفایر) به خط میباشد که فرمول آن به صورت زیر میباشد:

$$-\frac{f(x_0)}{||w||_2^2} * w$$

توضیحاتی که دادیم برای حالت affine بود و اگر بخواهیم به طور کلی یک الگوریتم برای کلسیفایر باینری معرفی کنیم به صورت زیر میباشد:

Algorithm 1 DeepFool for binary classifiers

input: Image $\boldsymbol{x}$, classifier f.
output: Perturbation $\hat{\boldsymbol{r}}$.
Initialize $\boldsymbol{x}_0 \leftarrow \boldsymbol{x}$, $i \leftarrow 0$.
while $\text{sign}(f(\boldsymbol{x}_i)) = \text{sign}(f(\boldsymbol{x}_0))$ **do**
    $\boldsymbol{r}_i \leftarrow -\frac{f(\boldsymbol{x}_i)}{\|\nabla f(x_i)\|_2^2} \nabla f(\boldsymbol{x}_i)$,
    $\boldsymbol{x}_{i+1} \leftarrow \boldsymbol{x}_i + \boldsymbol{r}_i$,
    $i \leftarrow i + 1$.
end while
return $\hat{\boldsymbol{r}} = \sum_i \boldsymbol{r}_i$.

توضیح الگوریتم:

1- الگوریتم ورودی x و کلسیفایر f را میگیرد.
2- خروجی کوچکترین perturbation مورد نیاز برای دسته بندی اشتباه تصویر میباشد.
3- مقدار دهی اولیه adver image با داده ورودی میباشد.
4- حلقه را شروع میکنیم و ادامه میدهیم تا تا زمانی که برچسب صحیح و برچسب perturbed image یکی باشد.
5- پروجکشن ورودی به نزدیکترین هایپرپلن را حساب میکنیم تا minimal perturbation به دست آید.
6- این perturbation را به تصویر اضافه و تست میکنیم.

7-8 – افزایش حلقه و اتمام حلقه

9- برگرداندن minimal perturbation

برای کلسیفایر های چند کلاسه برای هر کلاس یک هایپرپلن وجود دارد که یک کلاس را از بقیه کلاس ها جدا میکند و مبنی بر جایی که x در فضا دارد دسته بندی میشود. تمام کاری که این الگوریتم انجام میدهد این است که نزدیکترین هایپرپلن را پیدا میکند و این x را پروجکت میکند به آن هایپرپلن و آن را کمی فراتر از آن مرز میبرد تا دسته بندی اشتباه با minimal pertutbation رخ دهد.

Deepfool همچنین الگوریتمی را فراهم کرده است که بتواند تحمل پذیری adversarial را اندازه گیری کند.

منبع:

https://towardsdatascience.com/deepfool-a-simple-and-accurate-method-to-fool-deep-neural-networks-17e0d0910ac0