به نام خدا

گزارش تمرین عملی سوم درس

امنیت و حریم خصوصی در یادگیری ماشین

حمیدرضا امیرزاده

۴۰۱۲۰۶۹۹۹

الف

(۱)

برای آموزش استاندارد با محموعه دادگان Cifar-10 از هایپرپارامترها و تنظیمات زیر استفاده شد:

نرخ یادگیری: ۰,۰۱

تابع هزینه: Cross entropy loss

الگوریتم بهینه سازی: Stochastic gradient descent

تعداد ایپاک آموزش: ۱۰۰

بعد از آموزش شبکه ResNet روی این دادگان آموزش دقت مدل روی داده تست به میزان **۸۵,۰۵**٪ رسید.

```
Standard Accuracy of ResNet18 model on the 10000 test images: 85.05 %
```

(۲)

بعد از پیاده سازی حمله FGSM (به نوتبوک مراجعه شود)، دقت مدل آموزش دیده در مرحله قبل به ازای سه مقدار مختلف ε = 4/255, 8/255, 12/255 به صورت زیر مشاهده شد:

ε = 4/255، دقت : ۱۴,۰۱٪

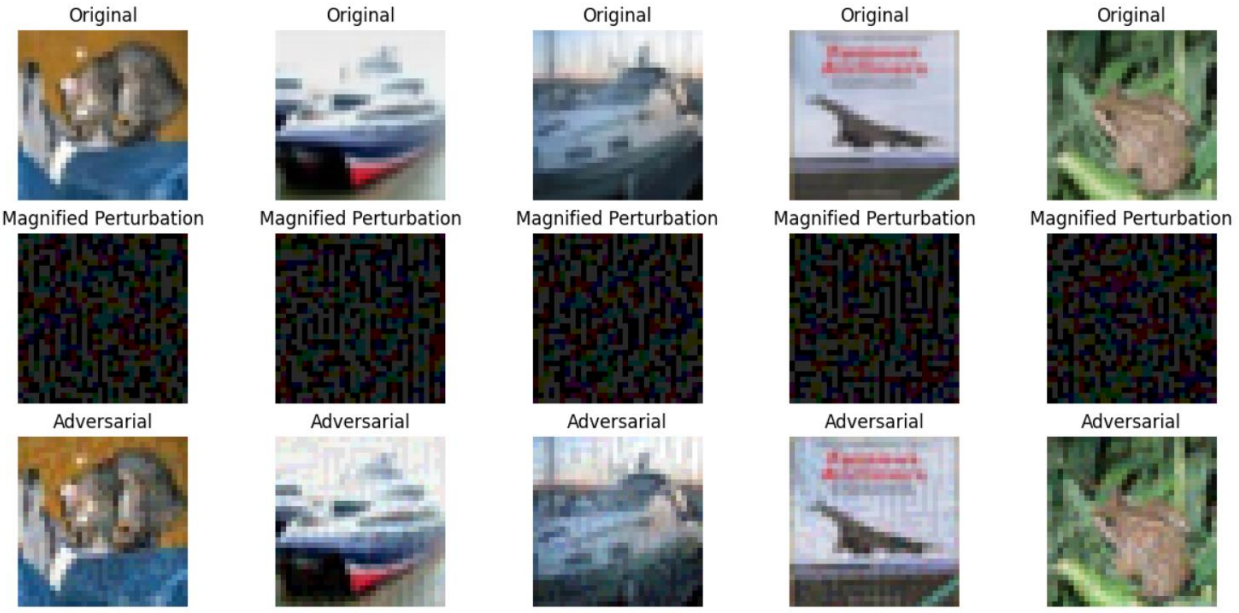
ε = 8/255، دقت : ۲,۷۹٪

ε = 12/255، دقت : ۱,۱۱٪

مشاهده می شود که با افزایش مقدار ε، حمله قوی تری انجام می شود و دقت مدل نیز کاهش بیشتری پیدا می کند. البته لازم به ذکر است که افزایش مقدار ε باعث محسوس شدن تغییر نمونه خصمانه می شود.

```
Accuracy of the model on the test images with epsilon 0.01568627450980392: 14.015%
Accuracy of the model on the test images with epsilon 0.03137254901960784: 2.787%
Accuracy of the model on the test images with epsilon 0.047058823529411764: 1.105%
```

(۳)

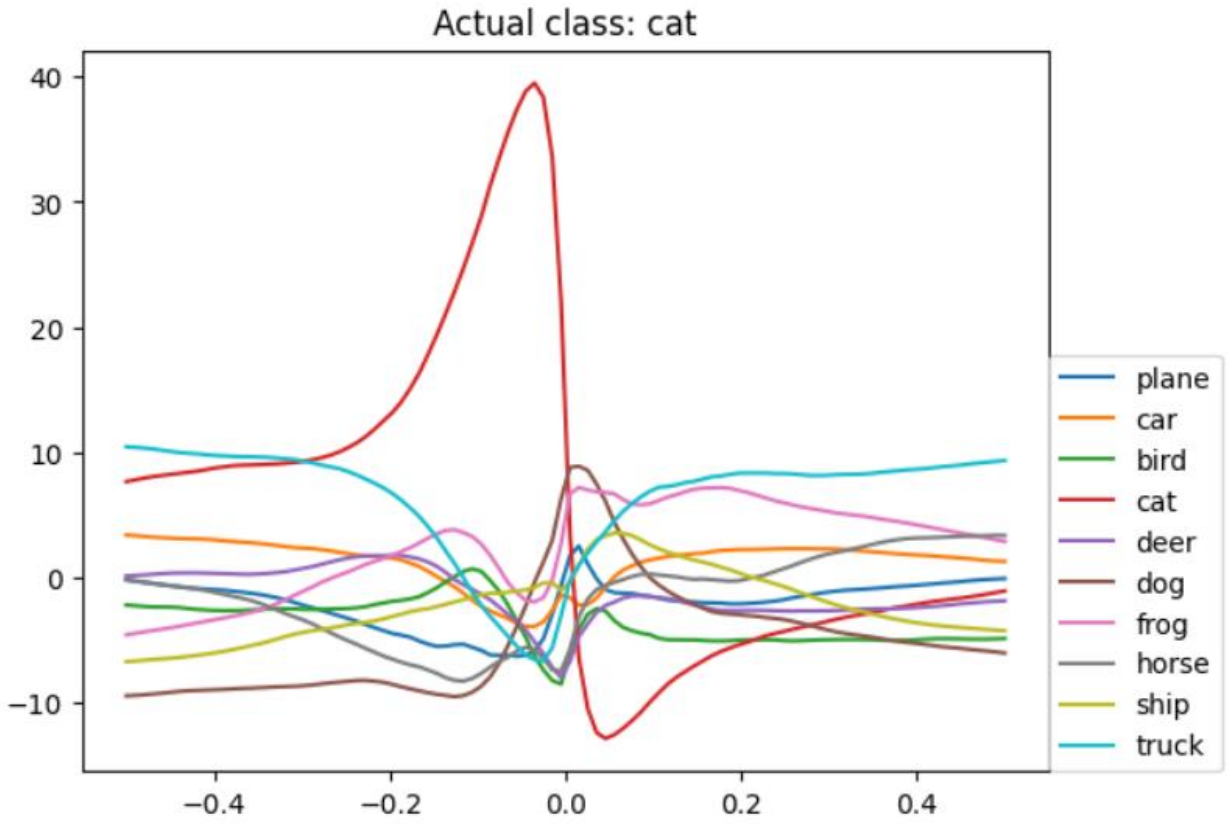



```
true label :  3
model prediction :  5
model prediction confidence :  0.7407677173614502
-----------------------------------
true label :  8
model prediction :  1
model prediction confidence :  0.9994565844535828
-----------------------------------
true label :  8
model prediction :  9
model prediction confidence :  0.5078034996986389
-----------------------------------
true label :  0
model prediction :  9
model prediction confidence :  0.9198201298713684
-----------------------------------
true label :  6
model prediction :  4
model prediction confidence :  0.9937684535980225
```

شکل بالا: نمایش ۵ نمونه اول خصمانه ساخته شده روی داده تست. ردیف اول داده تمیز و اصلی. ردیف دوم نویز به دست آمده توسط روش FGSM و با $\varepsilon = 8/255$. توجه شود برای اینکه نویز واضح تر دیده شود، ضربدر ۵ شده است. ردیف سوم نمونه خصمانه حاصل.

شکل پایین: به ترتیب برای از بالا به پایین متناظر است با تصاویر از چپ به راست در شکل اول. مشاهده می شود که در تمامی موارد مدل برای نمونه خصمانه پیش بینی اشتباه و با درصد اطمینان بالایی کرده است.

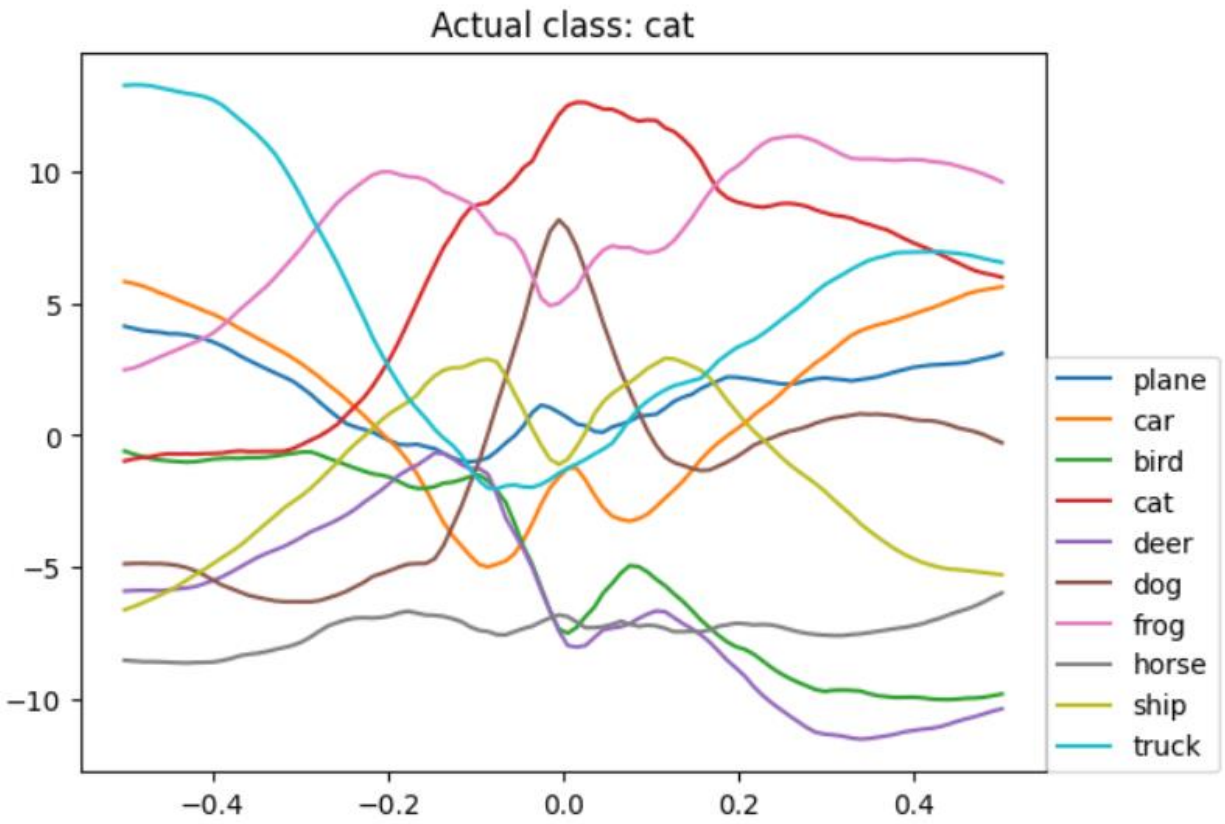
(۴)



این داده ای که برای بررسی انتخاب شده است متعلق به دسته ۳ یا Cat است. مشاهده می شود که وقتی $\varepsilon = 0$ است، لاجیت دسته ۳ بیشترین مقدار را دارد و دسته بند درست پیش بینی می کند. اما با حرکت کردن در جهت ساین گرادیان، بلافاصله از مرز تصمیم گیری عبور میکنیم و لاجیت دسته درست کاهش می یابد و منجر به تصمیم اشتباه دسته بند برای پیش بینی این داده می شود.

اگر در خلاف جهت ساین گرادیان حرکت کنیم تا مقادیر نسبتا خوبی از ε همچنان درون مرز تصمیم گیری می مانیم و پیش بینی دسته بند اشتباه نمی شود. اما بازهم به جایی می رسیم که مرز تصمیم گیری را در این جهت نیز رد کرده و مجددا تصمیم دسته بند اشتباه می شود.

در هر دو حالت تصمیم اشتباه دسته بند برای تصمیم گیری در اکثر مقادیر ε برابر دسته ۹ یا Truck است.



در این نمودار به جای آنکه در جهت ساین گرادیان حرکت کنیم، در جهت ساین یک تنسور تصادفی با توزیع U(-1,1) حرکت می کنیم. مشاهده می شود که در اینجا چون در یک جهت تصادفی حرکت کردیم، دیگر مانند حالت قبلی بلافاصله مرز تصمیم گیری را عبور نمی کنیم، زیرا جهت گرادیان بهترین جهتی بود که می توانستیم با حرکت به سمت آن تابع هزینه را به سمت بیشینه شدن سوق دهیم.

اما باز هم با این وجود، با بزرگ شدن مقدار ε مرز تصمیم گیری را رد می کنیم و لاجیت دسته درست کاهش می یابد و پیش بینی دسته بند عوض می شود.

ب

(۱)

برای انجام یادگیری خصمانه از هایپرپارامترها و تنظیمات زیر استفاده شد:

اپسیلون: ۸/۲۵۵

نرخ یادگیری: ۰,۰۱

تابع هزینه: Cross entropy loss

الگوریتم بهینه سازی: Stochastic gradient descent

برنامه ریز نرخ یادگیری: ضریب گاما برابر ۰,۹

تعداد ایپاک آموزش: ۱۰۰

بعد از آموزش خصمانه شبکه ResNet روی دادگان آموزش به دقت های زیر رسیدیم:

دقت روی دادگان تست تمیز و اصلی: ۷۴,۱۱٪

دقت خصمانه با حمله FGSM با ε = 8/255: ۷۷,۶٪

```
Accuracy of Adversarialy trained ResNet18 model on the 10000 clean test images: 74.11 %
Accuracy of Adversarialy trained ResNet18 model on the adversarial test images: 77.60141093474427 %
```

مشاهده می شود که آموزش خصمانه باعث کاهش دقت مدل روی دادگان تمیز و افزایش دقت مدل روی دادگان خصمانه می شود. بنابراین یک مصالحه بین عملکرد مدل و مقاومت آن وجود دارد.

(۲)

بعد از پیاده سازی حمله PGD (به نوتبوک مراجعه شود)، دقت مدل آموزش دیده به صورت استاندارد و مدل آموزش دیده به صورت خصمانه به ازای $\varepsilon = 8/255$ و دو مقدار k=2, 4 به صورت زیر مشاهده شد:

دقت مدل آموزش دیده به صورت استاندارد روی حمله PGD با k=2: ٪۰,۵۴

دقت مدل آموزش دیده به صورت استاندارد روی حمله PGD با k=4: ٪۰,۰۰

دقت مدل آموزش دیده به صورت خصمانه روی حمله PGD با k=2: ٪۸۷,۲۷

دقت مدل آموزش دیده به صورت خصمانه روی حمله PGD با k=4: ٪۶۴,۰۵

```
Accuracy of the standard model on the pgd adversarial test images with epsilon = 0.03137254901960784 and k = 2 is 0.54%
Accuracy of the standard model on the pgd adversarial test images with epsilon = 0.03137254901960784 and k = 4 is 0.00%
Accuracy of the adversarialy trained model on the pgd adversarial test images with epsilon = 0.03137254901960784 and k = 2 is 87.27%
Accuracy of the adversarialy trained model on the pgd adversarial test images with epsilon = 0.03137254901960784 and k = 4 is 64.05%
```

می بینیم که مدل استاندارد کاملا به حمله PGD ضعف دارد و این ضعف با افزایش تعداد گام های این حمله بیشتر هم می شود. اما مدل خصمانه تا حد بسیار خوبی نسبت به حمله PGD مقاومت پیدا کرده است.

(۳)

در قسمت های قبلی در این مورد بحث شد.

(۴)

بعد از اضافه کردن نویز گوسی با میانگین صفر و واریانس ۶/۲۵۵، دقت مدل استاندارد و مدل خصمانه به صورت زیر مشاهده شد:

دقت مدل استاندارد: ٪۳۱,۳

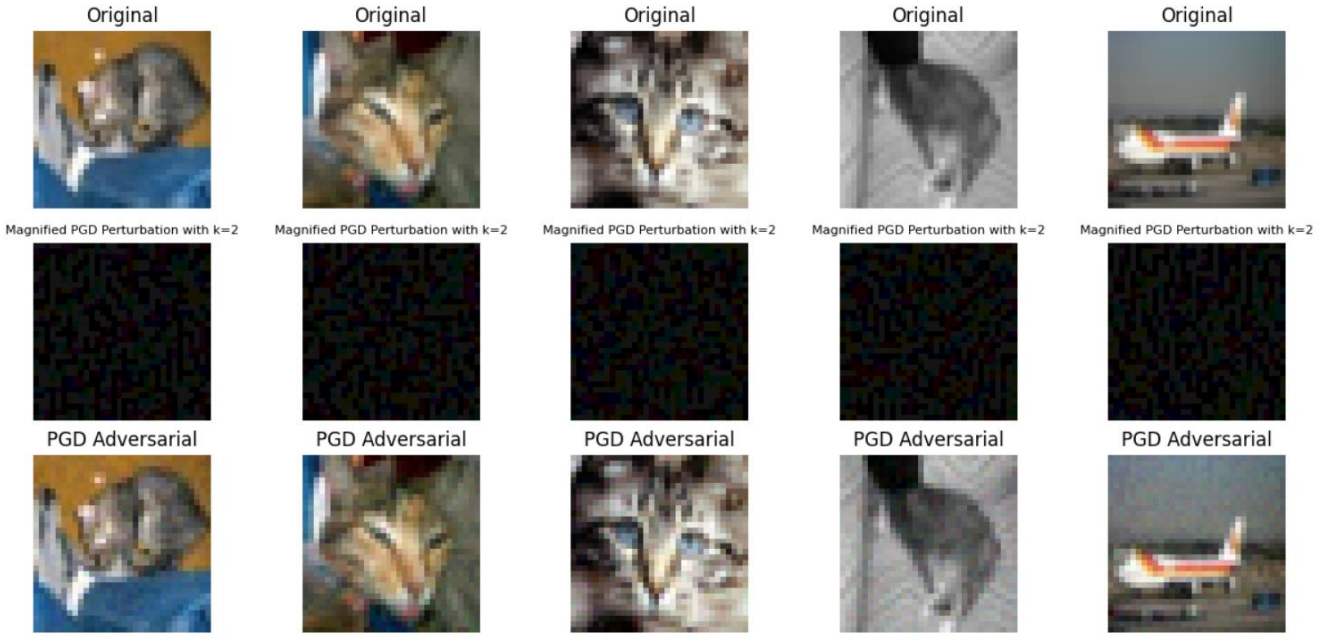
دقت مدل خصمانه: ٪۶۲,۳

```
Accuracy of Standard ResNet18 model on the 10000 test images with 0 mean and 6/255 variance gaussian noise: 31.3 %
```

```
Accuracy of Adversarialy trained ResNet18 model on the 10000 test images with 0 mean and 6/255 variance gaussian noise: 62.92 %
```

مجددا می بینیم که مدل خصمانه دقت و مقاومت بسیار بالاتری نسبت به مدل استاندارد دارد.

(۵)



نمایش ۵ نمونه اول خصمانه ساخته شده روی داده تست. ردیف اول داده تمیز و اصلی. ردیف دوم نویز به دست آمده توسط روش PGD و با $\varepsilon = 8/255$ و $k=2$. توجه شود برای اینکه نویز واضح تر دیده شود، ضربدر ۱۰ شده است. ردیف سوم نمونه خصمانه حاصل.