به نام خدا

تمرین سری سوم SPML ریحانه زهرابی ۴۰۰۲۰۱۱۱۲

گزارش سوال عملی:
در راستای بهبود دقت من ترنسفورم‌های مربوط به مجموعه داده را اندکی تغییر دادم.

الف) **مدل استاندارد:**

۱) دقت نهایی مدل یادگرفته شده روی مجموعه دادگان ارزیابی: ۹۲.۰۲

دقت نهایی مدل یادگرفته شده روی مجموعه دادگان آموزش: ۱۰۰

۲) دقت خصمانه مدل یادگرفته شده‌ی استاندارد در مقابل حمله‌ی FGSM :

[8] Evaluating FGSM Accuracy of Standard Trained Model

12/255	8/255	4/255	epsilon
43.58	47.42	51.63	adv accuracy

[9] Crafting Adversarial Examples with FGSM

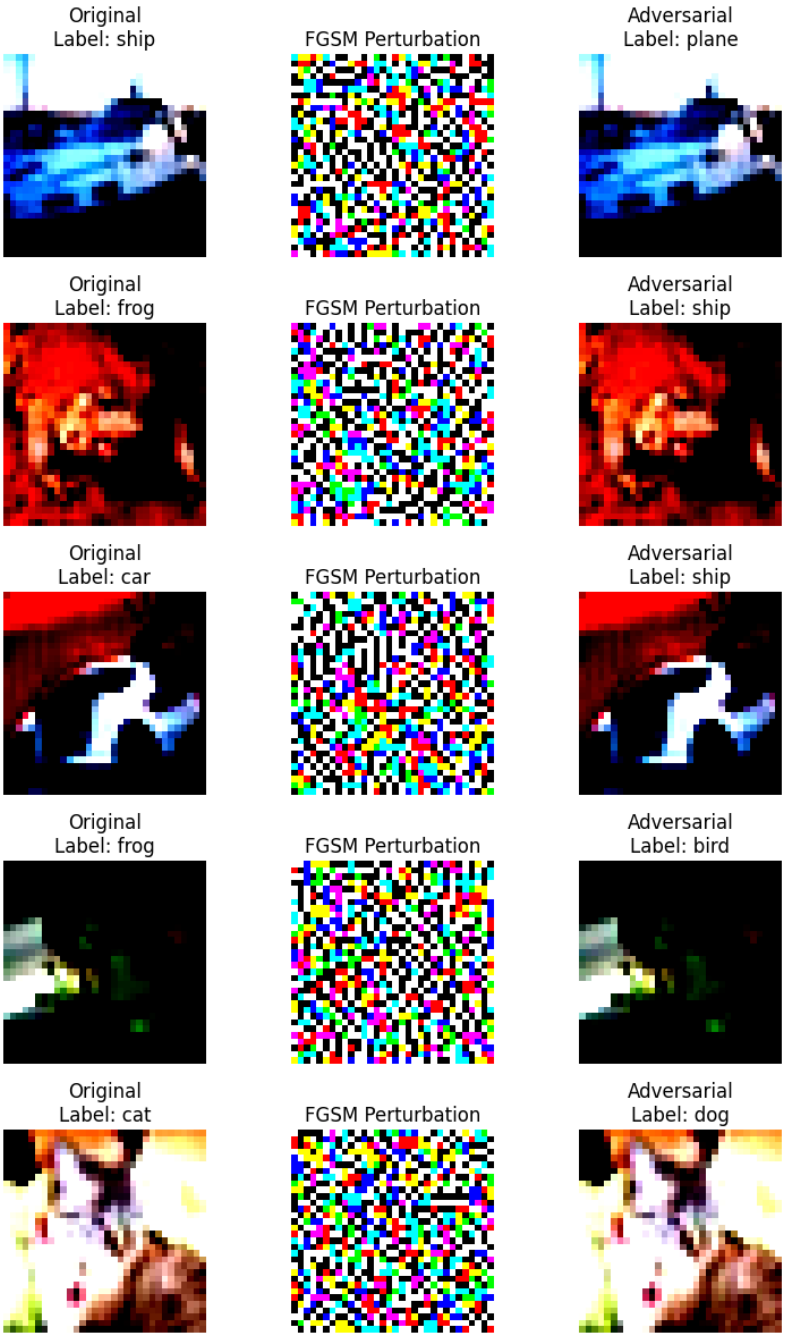
در این قسمت چون کلاس اتک را به گونه‌ای تغییر داده بودم که اتک‌های نا موفق نیز به مجموعه دادگان افزوده شوند آن‌هایی را چاپ کردم که اتک موفق بوده است.

```
Image 3
True Label: ship
Model Prediction: plane
Confidence: 0.63%

Image 6
True Label: frog
Model Prediction: ship
Confidence: 0.44%

Image 7
True Label: car
Model Prediction: ship
Confidence: 0.39%

Image 8
True Label: frog
Model Prediction: bird
Confidence: 0.88%

Image 9
```

```
True Label: cat
Model Prediction: dog
Confidence: 0.74%
```

(۳

[10] Plotting



ب) مدل خصمانه:

۱)

[11] Adversarial Training

```
Standard accuracy epsilon 0.03137254901960784: 10.750%
Adversarial accuracy epsilon 0.03137254901960784 : 85.38
```

در این قسمت به نظر می‌رسد به دلیل آموزش زیاد دقت روی داده‌های عادی کاهش چشمگیری داشته است.

۲و۳)

[13] Evaluating PGD Accuracy of Adversarially Trained Model and Standard Trained Model

```
Standard Trained Model:
PGD Accuracy (k=2): 40.73%
PGD Accuracy (k=4): 22.10%

Adversarially Trained Model:
PGD Accuracy (k=2): 81.92%
PGD Accuracy (k=4): 80.49%
```

همان‌گونه که واضح است مدلی که به صورت خصمانه آموزش دیده وقتی در برابر حمله‌ی pgd قرار می‌گیرد مقاومت بیشتری نسبت به مدلی دارد که به صورت عادی آموزش دیده است. از طرفی هرچه تعداد گام حمله بیشتر شده در واقع حمله قوی‌تر بوده و دقت نهایی در هر دو حالت کمتر شده است.

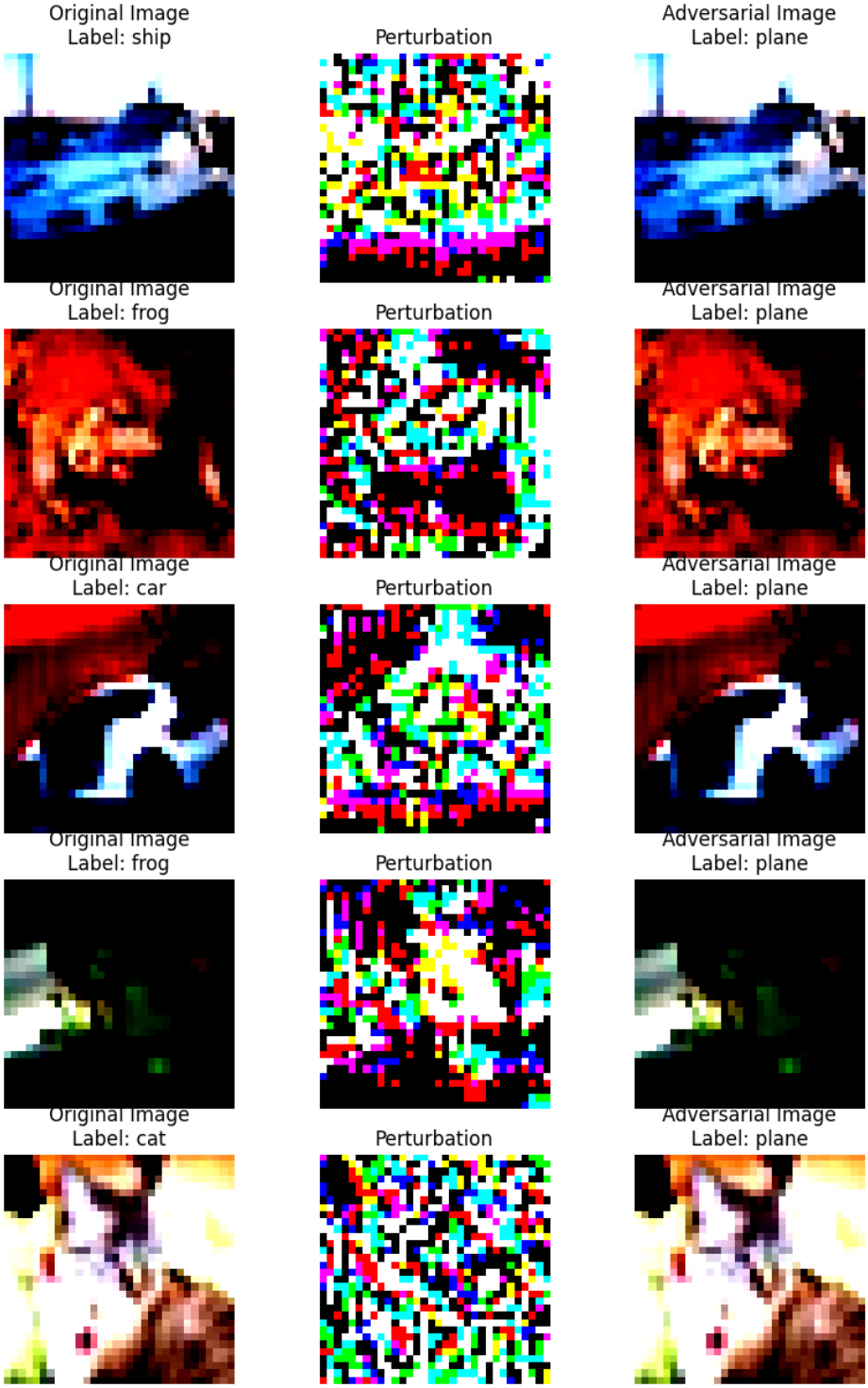
۴)

[15] Noisy Input Accuracy

```
Accuracy of the standard model on noisy images: 92.02%
Accuracy of the adversarial model on noisy images: 85.46%
```

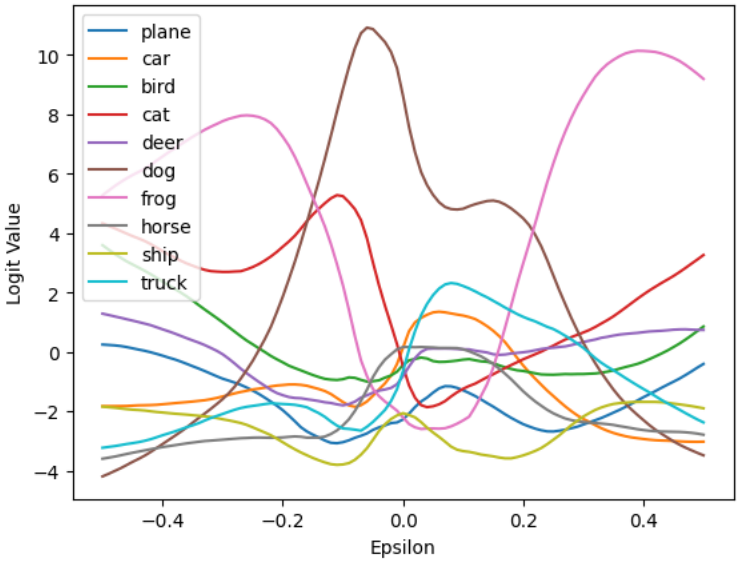
با توجه به نتایج می‌توان دید که وقتی نویز گاوسی به ورودی‌ها افزوده می‌شود دقت مدل‌ها کاهش زیادی نداشته است. اما وقتی از نمونه‌های خصمانه استفاده می‌شود دقت مدل‌ها به طور چشمگیری کاهش می‌یابد. بنابراین آشفتگی‌های خصمانه موفقیت بیشتری نسبت به نویزهای تصادفی در شکست دادن مدل‌ها دارند.

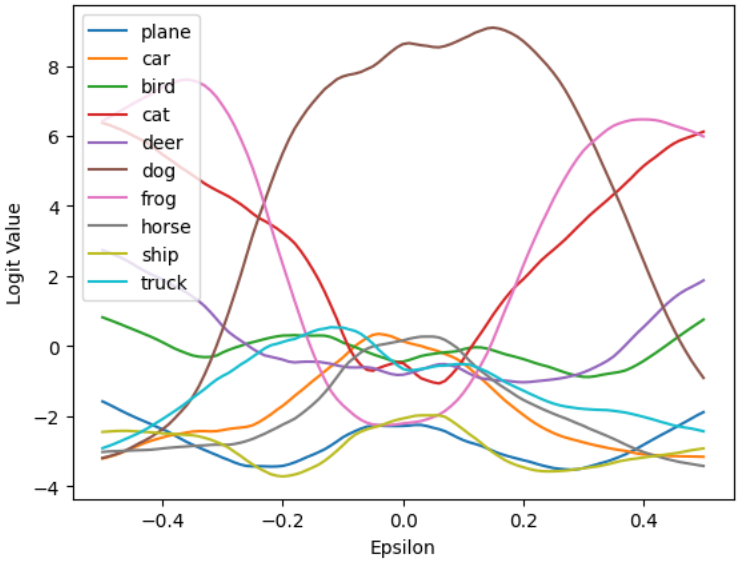
[14] Plotting


Original Image
Label: ship
Perturbation
Adversarial Image
Label: plane
Original Image
Label: frog
Perturbation
Adversarial Image
Label: plane
Original Image
Label: car
Perturbation
Adversarial Image
Label: plane
Original Image
Label: frog
Perturbation
Adversarial Image
Label: plane
Original Image
Label: cat
Perturbation
Adversarial Image
Label: plane

[16] Logit VS Epsilon Test

```
Predicted label: dog, True label: dog
```





نمودار اول تأثیر اغتشاشات در جهت گرادیان لاس را با توجه به تصویر ورودی نشان می دهد. همانطور که می بینیم، اغتشاشات کوچک روی مقادیر لاجیت تاثیر چندانی نمی گذارد، اما با افزایش بزرگی اغتشاشات، مقادیر لاجیت به طور قابل توجهی شروع به تغییر می کنند. این نشان می دهد که مدل به تغییرات کوچک در ورودی حساس است و نمونه های خصمانه را می توان به راحتی با اعمال اغتشاشات کوچک در ورودی ایجاد کرد.

نمودار دوم تأثیر اغتشاشات را در جهت تصادفی نشان می دهد. همانطور که می بینیم، مقادیر لاجیت در مقایسه با نمودار اول بسیار کمتر تغییر می کند، که نشان می دهد اغتشاشات در جهت تصادفی در ایجاد نمونه های خصمانه موثر نیستند. این نشان می‌دهد که مدل نسبت به اغتشاش‌ها در جهت‌های تصادفی حساس نیست و نمونه‌های متخاصم ایجاد شده در این راه ممکن است به همان اندازه مؤثر نباشند.
به طور کلی با ردیابی مقادیر مختلف اپسیلون، می‌توانیم ببینیم که نمونه‌های متخاصم تقریباً برای هر مقدار به اندازه کافی بزرگ به‌طور قابل اعتمادی رخ می‌دهند، مشروط بر اینکه در جهت درست حرکت کرده باشیم.