1.模型结构

generator：128+20—256—128

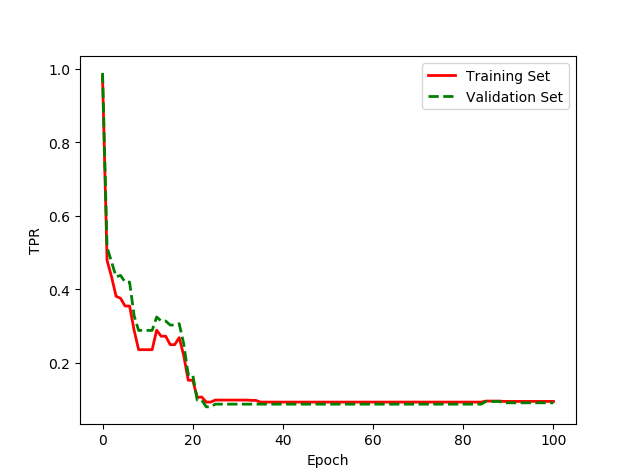
subsitite detector：128—256—1

maximum epoch：100

2.实验

（1）True positive rate (in percentage) on original samples and adversarial examples when MalGAN and the black-box detector are trained on the same training set. “Adver.” represents adversarial examples

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Training Set | | Test Set | |
|  | Original | Adver. | Original | Adver. |
| RF | 98.32 | 2.73 | 98.87 | 3.56 |
| LR | 98.53 | 10.32 | 99.27 | 11.67 |
| DT | 0 | 0 | 100 | 0 |
| SVM | 98.08 | 17.36 | 97.44 | 14.23 |
| MLP | 98.08 | 0.71 | 98.17 | 0 |
| VOTE | 98.72 | 12.13 | 96.71 | 13.51 |



The change of the true positive rate on the training set and the validation set over time. MLP is used as the black-box detector here. The vertical axis represents the true positive rate while the horizontal axis represents epoch.

（2）True positive rate (in percentage) on original samples and adversarial examples when MalGAN and the black-box detector are trained on different training sets. “Adver.” represents adversarial examples.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Training Set | | Test Set | |
|  | Original | Adver. | Original | Adver. |
| RF | 96.88 | 4.64 | 95.27 | 7.93 |
| LR | 94.07 | 12.53 | 96.84 | 15.24 |
| DT | 0 | 0 | 100 | 0 |
| SVM | 98.17 | 18.84 | 96.71 | 14.26 |
| MLP | 97.98 | 1.52 | 97.81 | 0.63 |
| VOTE | 98.35 | 13.81 | 98.9 | 14.15 |

（3）True positive rate (in percentage) on the adversarial examples after the black-box detector is retrained

|  |  |  |
| --- | --- | --- |
|  | Before Retraining | After Retraining |
| Training Set | 100 | 7.13 |
| Test Set | 100 | 12.31 |