Preferred Networks インターン選考 2018
レポート

課題 2：

予測モデルの正解率を確認するために、次の関数でプログラムを実行します。

▼リスト 1

```
$> ./attack -a none -t all
Success : 129 / 154
Accuracy : 83%
```

「-a」は攻撃方法を選ぶための関数です。ここでは摂動を追加しないので、「none」にします。「-t」は予測の対象を選ぶための関数です。全ての画像を予測して正解率を求めたい場合、「all」にします。予測モデルの正解率は 83%になります。

課題 3：

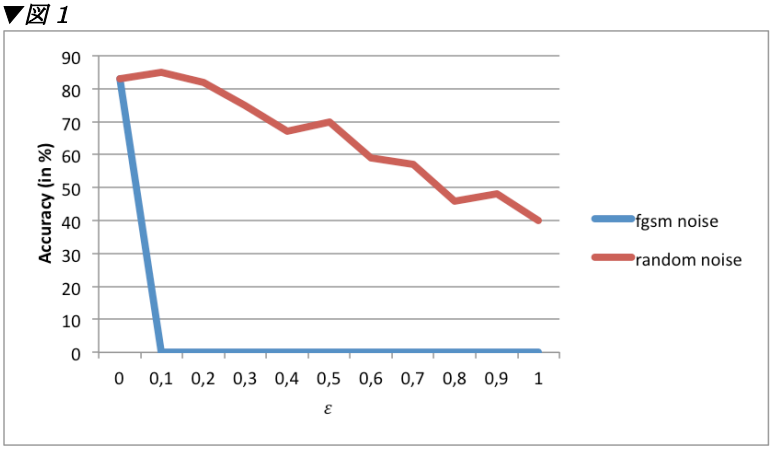
次のコマンドを入力すると、FGSM の摂動の攻撃方法の正解率とランダムの摂動の攻撃方法を比較できます。

▼リスト 1

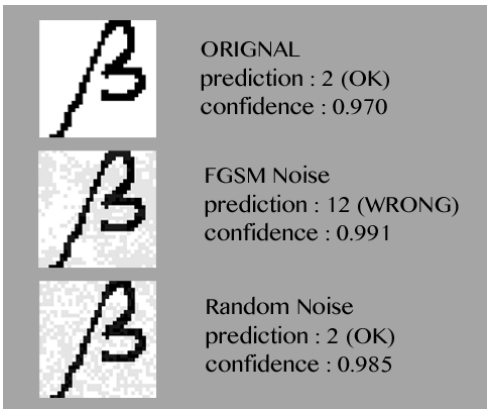
```
$> ./attack -a fgsm -t all -e 0.1
Before attack :
Success : 129 / 154
Accuracy : 83%
After attack (without binarize protection) :
Success : 1 / 154
Accuracy : 0%
After attack (with binarize protection) :
Success : 129 / 154
Accuracy : 83%
$> ./attack -a random -t all -e 0.1
Before attack :
Success : 129 / 154
Accuracy : 83%
After attack (without binarize protection) :
Success : 130 / 154
Accuracy : 84%
After attack (with binarize protection) :
Success : 129 / 154
Accuracy : 83%
```

「-e」はパラメータεを設定するための関数です。上記のように、FGSM の摂動の攻撃方法で正解率が 0%に落ちるが、ランダムの摂動の攻撃方法で正解率が 84%に増加します。「ε」による正解率の推移が次の図 1 で見られます。

▼図 1



▼図 2



FGSM の摂動の攻撃方法の能率を図 1 と図 2 で確認できます。ただし、入力をモノクロ化すれば、この方法は無効になります。

課題 4：

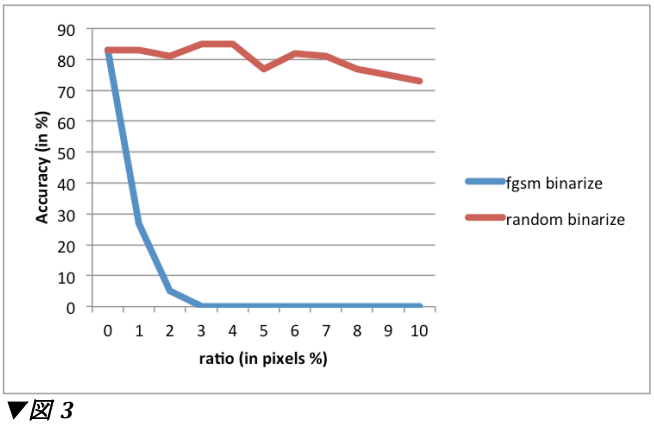
モノクロ化に対処するために、画像の全てのピクセルを少しずつ変えるのではなく、少しの割合のピクセルを反転（白を黒に、黒を白に）します。この反転するピクセル は、Gradient Descent を用いて入力の傾斜度を計算し、ピクセルを傾斜度によって並べ替え、傾斜度が一番大きいピクセルからいくつかのピクセルが選ばれます（図 3 参照）。次のコマンドでこの攻撃を実行できます。

▼リスト 1

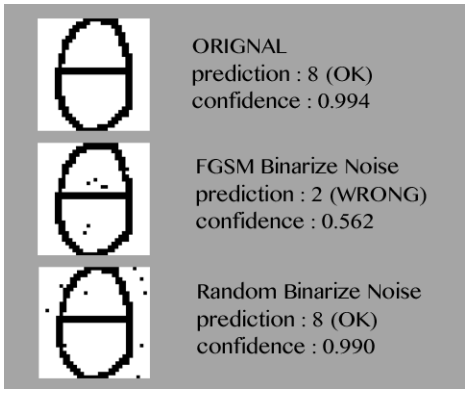
```
$> ./attack -a fgsm-binarize -t all -r 0.01
Before attack :
Success : 129 / 154
Accuracy : 83%
After attack (without binarize protection) :
Success : 43 / 154
Accuracy : 27%
After attack (with binarize protection) :
Success : 43 / 154
Accuracy : 27%
$> ./attack -a random-binarize -t all -r 0.01
Before attack :
Success : 129 / 154
Accuracy : 83%
After attack (without binarize protection) :
Success : 130 / 154
Accuracy : 84%
After attack (with binarize protection) :
Success : 130 / 154
Accuracy : 84%
```

「-r」はピクセルの割合を設定するための関数です。0.01 で 1 %のピクセルを反転します。リスト 1 で表示されるように、この方法で、モノクロ化のモデルの正解率と、モノクロ化がないモデルの正解率が同じになり、27%に減ります。反転するピクセルをランダムに選ぶと、正確度が 84%に増えます。ピクセルの割合による正解率の推移が図 1 で見られます。

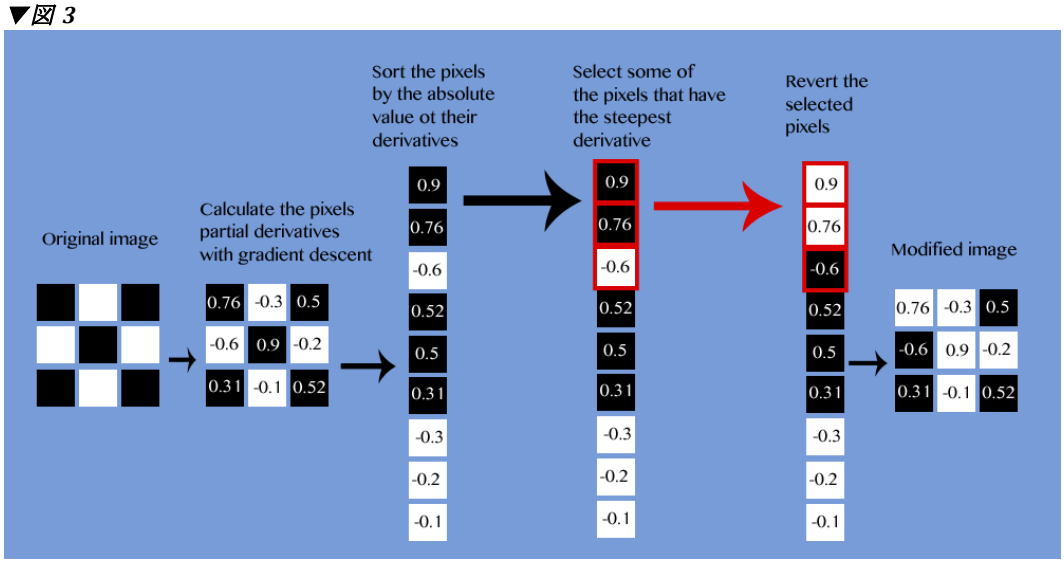
▼図 1



▼図 2



▼図 3



つまり、傾斜度の大きさによって少しの割合のピクセル を反転することで、入力をモノクロ化してしまうモデルを攻撃できました。

Merlin Dugot