課題 2
予測モデルの正答率は 0.838

課題 3

元の画像の画素をベクトル化したものに摂動を加えたあと、32×32 個の画素の最大値を 255、最小値を 0 とするような線形変換をした（端数は四捨五入して整数値にする）。
元の画像にランダムな摂動を加えた画像（ベースライン）では正答率は 0.81 であるのに対し、FGSM に基づく摂動を加えた画像では正答率は 0.013 となった。

また、ε0 は小さいため、摂動を加えた画像の（人間から見た）外見はあまり違わず、元通りの文字として認識できる。

図 1　左から順に元の画像・ランダムな摂動を加えた画像・FGSM に基づく摂動を加えた画像

課題 4（2 値化による防御とそれに対する攻撃）

4.1　画像を 2 値化してから分類することにより、FGSM から防御する
入力した番号の pgm ファイルを読み込み、画素の値を左上から右方向に読み込み、1024 次元のベクトルとする関数 pgm_to_vec_binary(pgm ファイルの番号)を実装した。
ただし、元の画像の画素値が 129 以上のピクセルは全て 255（白）、画素値 128 以下ピクセルは全ては 0（黒）とするという 2 値化処理をしている。
課題 3 で摂動（ノイズ）を与えた画素の値は小さくても 230（白に近い灰色）程度なので、ニューラルネットワーク予測モデルに入力する前に 255（白）になるので FGSM の攻撃は防御できる。

4.2　2 値化による防御をされても突破できる攻撃方法についての考察

以下の目標を設定して攻撃する。
1.元の画像のいくつかのピクセルの画素値を 122（実際には2値化のスレッショルドに応じて「黒」にされるグレーであれば何でも良い）にして、ニューラルネットワーク予測モデルに分類を誤らせる。
2.1.の処理をした画像を人間が目で見た時には元の文字と認識し、かつ可能な限り 1.の攻撃的な処理をしたことに気づかないようにする。

そこで、課題 3 と同じ方法で画像ごとにクロスエントロピー誤差関数を各画素の値で偏微分したもの（∇xL）を求め、画素値が下がるときにクロスエントロピー誤差を大きくする（＝予測モデルが間違いやすくなる）ようなピクセルを探し当てる。私のレポートでは、∇xL に対して閾値を設け、それを下回る偏微分係数を持つピクセルの画素値が 2 値化後 0（黒）になるような方法をとることにした。

この攻撃の結果、クロスエントロピー関数の偏微分係数の閾値を変えた場合のニューラルネットワークの正答率は以下の表の通りである。

閾値	-0.0002	-0.0004	-0.0008	-0.0012
画像を 2 値化した後で予測した時の正答率	0.11	0.15	0.20	0.25
画像を 2 値化せず、そのまま予測した時の正答率	0.13	0.16	0.23	0.27
画像ファイルの置き場所	problem4_1/	problem4_2/	problem4_3/	problem4_4/

また、画像は以下のような見た目になった。

図 2　閾値を-0.0012 にした時の摂動を加えたあとの画像。正しく読めたのは左から数えて 1,2,5,7 番目。

4.3　より良い攻撃の実装

4.2 の方法では、画像によってはあまりにたくさんの摂動が加わってしまい、人間の目をごまかせなくなってしまうことがある（図 2 の右 3 つなど）。そこで、この節ではクロスエントロピー誤差関数の偏微分係数が小さいピクセル n（=100,50,30,15)個を、そこの画素値が 2 値化後に 255（白）であるならば 122（2 値化後黒になる）にするという手法を取る。この方法ならば、摂動により変化するピクセルの数を制限することができるので人間の目で文字を正しく読むことを阻害しないだろう。

この攻撃の結果は以下の通りである。

摂動を与えるピクセル数	100	50	30	15
画像を 2 値化した後で予測した時の正答率	0.0	0.0	0.045	0.299
画像ファイルの置き場所	problem4_5/	problem4_6/	problem4_7/	problem4_8/

また、画像は以下のような見た目になった。

図 3　摂動を与えるピクセル数を 50 個にした時（左 5 つ）及び 15 個にした時（右 5 つ）に摂動を加えたあとの画像の例。

表及び図 3 に示した通り、確かに人間の見た目にはほとんど変化がない一方で、ニューラルネットを完全に”騙す”ことに成功した。摂動を与えるピクセル数を調節することで、うまく人間のチェックをすり抜けてかつニューラルネットワークに誤答させることができるだろう。

感想

課題 4.3 で行った攻撃は我ながらかなりの威力であったと考えている。特に図 3 の右側の画像は、人間の目には印刷のカスレにしか見えない程度の摂動でニューラルネットワークに文字を認識できなくしていると言えるだろう。

また、2 値化するという防御策によってかえって攻撃者が与えた摂動の効果を増大させ、クロスエントロピー誤差関数を大きくしてしまっているのが面白いと思った。