CS612 Assignment 1

Kasun Gamlath

September 6, 2025

Exercise 1

Assuming the reference image R is a black image with all pixel values 0, i.e., R = [0, 0, ..., 0] (784 zeros for a 28x28 image).

$L_1 - norm$ with a threshold (inclusive) of 2

Condition: $\sum |r_i - c_i| <= 2$
Since $r_i = 0$, this becomes $\sum |c_i| <= 2$
As pixel values are non-negative, $\sum c_i <= 2$.
This means the sum of all pixel values in the candidate image C must be 0, 1, or 2.

- Case 1: $\sum c_i = 0$

  This means all c_i must be 0.

  Only 1 image: [0, 0, ..., 0] (which is identical to R).

- Case 1: $\sum c_i = 1$

  One pixel c_k is 1, and all other pixels are 0.

  There are 784 possible positions for this 1.

  So, 784 images. (e.g., [1, 0, ..., 0], [0, 1, ..., 0], etc.)

- Case 3: $\sum c_i = 2$

  - Subcase 3a: One pixel c_k is 2, and all other pixels are 0.
    There are 784 possible positions for this 2.
    So, 784 images.
  - Subcase 3b: Two pixels c_k and $c_j (k \neq j)$ are 1, and all other pixels are 0.
    This is "784 choose 2" (combinations). C(784, 2) = 784 * 783 / 2 = 306912.
    So, 306912 images.

Total Images for $L_1 =< 2 = 1 + 784 + 784 + 306912 = 308481$

$L_2 - norm$ with a threshold (inclusive) of 2

Condition: $\sqrt{\sum (r_i - c_i)^2} <= 2$

- Case 1: $\sum c_i = 0$

  All c_i must be 0.

  Only 1 image: [0, 0, ..., 0].

- Case 2: $\sum c_i = 1$

  One pixel c_k is 1, all others 0. $1^2 = 1$.

  There are 784 positions for this 1.

  So, 784 images.

- Case 3: $\sum c_i = 2$

  Two pixels c_k, c_j are 1, all others 0. $1^2 + 1^2 = 2$.

  There are C(784, 2) ways to choose these two positions. C(784, 2) = 306912.

  So, 306912 images.

- Case 4: $\sum c_i = 3$

  Three pixels c_k, c_j, c_m are 1, all others 0. $1^2 + 1^2 + 1^2 = 3$.

  There are C(784, 3) ways. C(784, 3) = 784 * 783 * 782 / (3 * 2 * 1) = 80119296.

  So, 80119296 images.

- Case 5: $\sum c_i = 4$
  - Subcase 5a: One pixel $c_k is 2$, all others 0. $2^2 = 4$.
    There are 784 positions for this 2.
    So, 784 images.
  - Subcase 5b: Four pixels c_k, c_j, c_m, c_n are 1, all others 0. $1^2 + 1^2 + 1^2 + 1^2 = 4$.
    There are C(784, 4) ways. C(784, 4) = 784 * 783 * 782 * 781 / (4 * 3 * 2 * 1) = 15683222004.
    So, 15683222004 images.

  Total Images for $L_2 =< 2 = 1+784+306912+80119296+784+15683222004 \approx 15.76 Billion$

$L_\infty - norm$ with a threshold (inclusive) of 2

Condition: $\max(|r_i - c_i|) =< 2$ Since $r_i = 0$, this becomes $\max(c_i) =< 2$.

This means that every single pixel c_i in the candidate image C must have a value between 0 and 2 (inclusive).

For each of the 784 pixels, it can independently take on one of 3 values: 0, 1, or 2.

Total Images for $L_\infty =< 2 = 3^{784}$

Exercise 2

eps	untargeted	targeted
0.2	5/5	1/5
0.1	4/5	0/5
0.05	1/5	0/5
0.01	0/5	0/5

Exercise 3

modification	Number of labels changed
(-0.01,0.01)	1/20
(-0.05,0.05)	6/20
(-0.1,0.1)	11/20

Conclusion: When the noise gets large, more adversarial samples lose their adversarial label.

Exercise 4

eps	vanila	adversarial trained
0.2	20/20	19/20
0.1	12/20	9/20
0.05	4/20	3/20
0.02	0/20	0/20

Observation: Model with adversarial training has fewer labels changed.

Conclusion: Model with adversarial training is more robust.

Exercise 5

	vanila	robust
training time (sec)	8.2	8.7
accuracy (%)	93	94.22
success rate of FGSM attack (%)	97	69.6