Midterm Score Review

• Midterm Grade on SIO is not final grade
• 2 Homeworks (36 pt)
• Midterm Participation Grade (5 pt)
  • 1 pt if seen in class regularly
  • Other 4 pts distributed across:
    • In-class Participation
    • After-class Participation
    • Piazza Participation
## Midterm Stats

<table>
<thead>
<tr>
<th></th>
<th>HW1</th>
<th>HW2</th>
<th>Midterm Participation</th>
<th>Midterm Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
<td>12.20</td>
</tr>
<tr>
<td>Max</td>
<td>18.00</td>
<td>19.00</td>
<td>5.00</td>
<td>101.83</td>
</tr>
<tr>
<td>Median</td>
<td>17.50</td>
<td>17.50</td>
<td>3.50</td>
<td>90.85</td>
</tr>
<tr>
<td>Standard dev</td>
<td>2.98</td>
<td>3.00</td>
<td>1.52</td>
<td>13.33</td>
</tr>
<tr>
<td>Average</td>
<td>16.43</td>
<td>16.68</td>
<td>3.20</td>
<td>88.55</td>
</tr>
</tbody>
</table>
Grade Distribution

Count of Midterm Grade

<table>
<thead>
<tr>
<th>Grade</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>15</td>
</tr>
<tr>
<td>A-</td>
<td>25</td>
</tr>
<tr>
<td>B+</td>
<td>17</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
</tr>
<tr>
<td>B-</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
</tr>
<tr>
<td>R</td>
<td>1</td>
</tr>
</tbody>
</table>
Grade Percentage Distribution

Histogram of Midterm Percentage

Midterm Percentage
HW4 Part I

18739
Caleb Kaiji Lu
HW4 Logistics

- 2 parts on adversarial models in DL
  - Targeted/evasion attack
  - Membership inference attack
- Part I to be released today
Szegedy et al. 2014, *Intriguing properties of neural networks*

“We describe a way to traverse the manifold represented by the network in an efficient way and finding adversarial examples in the input space”

\[
\begin{align*}
\text{Minimize } & \|r\|_2 \text{ subject to:} \\
& 1. f(x + r) = l \\
& 2. x + r \in [0, 1]^m
\end{align*}
\]

Minimize to make “inconspicuous”

Attacker’s main objective

Still a valid input
Optimization Problem

- Form 1:
  
  \[
  \text{Minimize } \|r\|_2 \text{ subject to:}
  \]
  
  1. \( f(x + r) = l \)
  
  2. \( x + r \in [0, 1]^m \)

- Form 2:
  
  Minimize \( c|r| + \text{loss}_f(x + r, l) \) subject to \( x + r \in [0, 1]^m \)
Implementation in tensorflow

- **Operation 1:**
  - x is the adversarial image (tf.Variable) to be learned
  - GradientDescentOptimizer that minimize loss(f(x), l)

- **Operation 2:**
  - With a small c, clip x at each time step t so that it is:
    - Between [x_o+c, x_o-c], where x_o original image
    - Between [0,1] for each dimension of x

- **Stop whenever the prediction is flipped to the target class**

- **We will provide with main function that:**
  - Checks if the prediction is flipped
  - Calculates a target distortion so that it is within a certain range
Example Adversarial Images