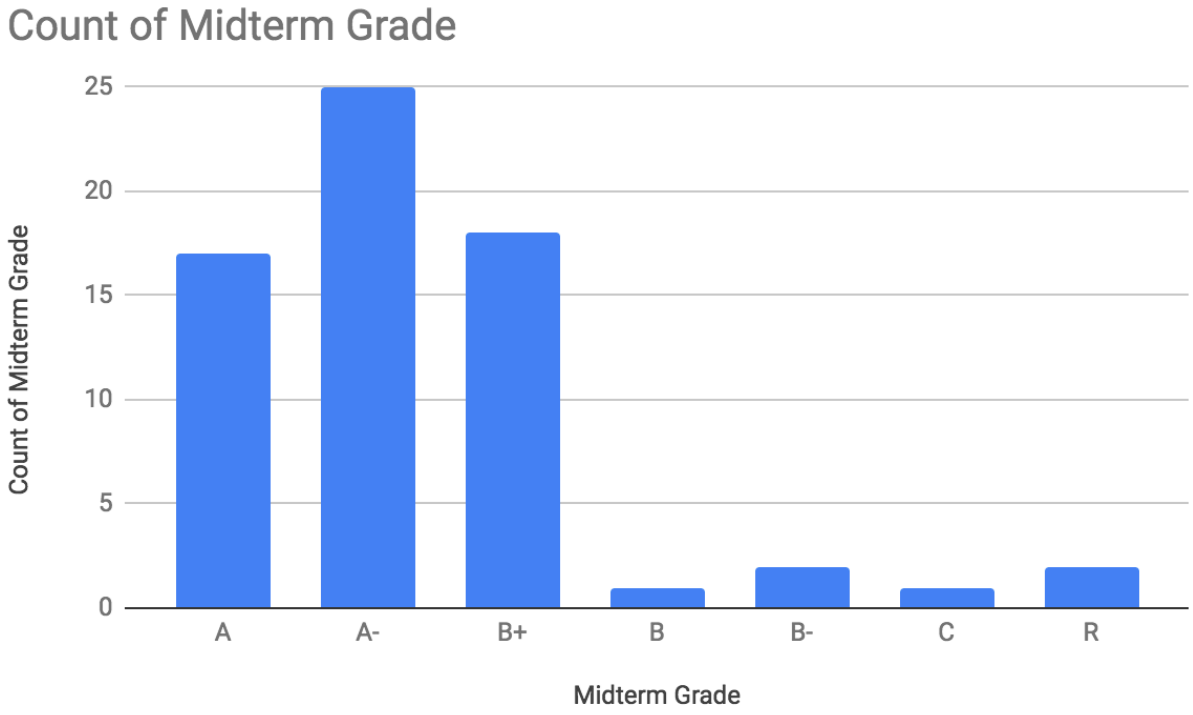
Midterm Score Review

- 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

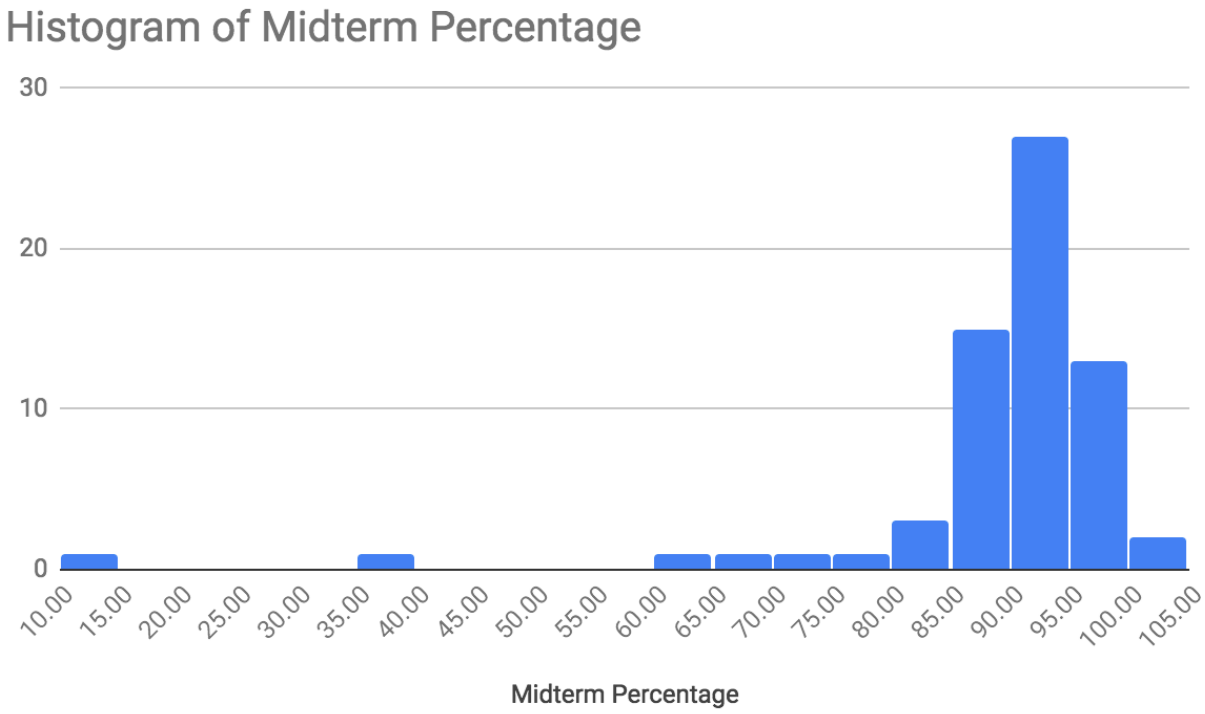
Midterm Stats

	HW1	HW2	Midterm Participation	Midterm Percentage
Min	1.00	0.00	0.00	12.20
Max	18.00	19.00	5.00	101.83
Median	17.50	17.50	3.50	90.85
Standard dev	2.98	3.00	1.52	13.33
Average	16.43	16.68	3.20	88.55

Grade Distribution



Grade Percentage Distribution



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

Review: Targeted Attack In Deep Learning

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*"

Minimize $\|r\|_2$ subject to:

1. $f(x+r) = l$
2. $x + r \in [0,1]^m$

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:

$$\text{Minimize } c|r| + \text{loss}_f(x+r, l) \text{ subject to } x + r \in [0,1]^m$$

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