Neural Networks Part 2 Stanford CS231N

Name: Eli Andrew

- **Activation Functions**
  - **Sigmoid**
    - $\sigma(x) = \frac{1}{1+e^{-x}}$
    - Saturated neurons (output close to 0 or 1) "kill" the gradients.
    - Sigmoid outputs are not zero-centered. If input to neuron is always all positive then gradient is either all positive or all negative. This leads to bad gradient updates.
    - **This is why you want zero-mean data**
    - e^{-x} is expensive to compute (compared to other options)
  - **Tanh**
    - Attempts to solve the issue of being zero-centered.
    - Still kills the gradients when saturated.
  - **ReLU**
    - Computes $f(x) = max(0, x)$
    - Neuron does not saturate in positive region (doesn't kill gradients).
    - Very computationally efficient
    - Converges must faster than other options.
    - Does not have zero-centered data
    - Kills gradient when less than 0.
    - Dead ReLU problem happens when neuron gets knocked off "data-manifold" and then can no longer update.
    - Initialize the ReLU units with slightly positive bias (0.01) to try and avoid the dead ReLU issue.
  - **Leaky-ReLU**
    - Computes $f(x) = max(0.01x, x)$
    - Does not die like regular ReLU
- **Data Pre-processing**
  - Zero-center data (subtract mean from every feature)
  - Normalize data (not as common for images)
- **Weight Initialization**
  - Zero weight initialization doesn't break symmetry so all neurons update the same.