Activation functions in neural networks

Sigmoid $f(x)=\frac{1}{1+e^{-x}}$

Description

- squashes numbers to range [0, 1]
  high values near 1, high negative values near 0
- has a nice interpretation of saturating the "firing rate" of a neuron

Problems

- saturated neurons "kill" the gradient
  high positive and high negative values generate ~0 gradients (flat slope)
- sigmoid outputs are not zero-centered (inneficient gradient updates)
- the exponential function is computationally expensive

Tanh $f(x)=tanh(x)$

Description

- squashes numbers to range [-1, 1]
  high values near 1, high negative values near -1
- outputs are zero-centered

Problems

- saturated neurons "kill" the gradient

ReLU (REctified Linear Unit) $f(x)=max(0,x)$

Description

- does not saturate in the positive region
- very computationally efficient
- converges much faster than sigmoid/tanh in practice

Problems

- not zero-centered output
- saturated neurons in the negative region
- dead ReLUs will never activate and therefore will never update

Leaky ReLU $f(x)=max(0.1x,x)$

Description

- does not saturate
- computationally efficien
- converges faster than sigmoid/tanh in practice
- will not die