Mish: A Self Regularized Non-Monotonic Neural Activation Function

Gobind Puniani
MATH 190: Fall 2019
11/14/19

Activation Functions

- Activation functions introduce non-linearity into the network
  - Vital for training and testing
- Only a select few activation functions relied upon:
  - ReLU (Rectified Linear Unit): $f(x) = \max(0, x)$
  - Swish: $f(x) = x \cdot \text{sigmoid}(x)$
  - TanH (Hyperbolic Tangent)
  - Sigmoid
  - Leaky ReLU

ReLU and Swish

- ReLU and Swish are the most popular activation functions
- ReLU
  - Standard/default activation
  - Simple Implementation
  - Virtually unchallenged
- Swish
  - Unbounded above and bounded below

Mish

- Mish function:
  $f(x) = x \cdot \tanh(\text{softplus}(x)) = x \cdot \tanh(\ln(1 + e^x))$
- Benefits of Mish:
  - Similar to Swish
  - Better performance
  - Easy to implement

Properties of Mish

- Self-gating: scalar input is provided to the gate
- Training factors:
  - Bounded below and unbounded above
  - Smooth
  - Non-monotonic
  - Other factors may help (hard to determine)
- Implemented on any standard deep learning framework
  - Lower learning rate recommended

Experimental Results

- Compared to Swish and ReLU:
  - Mish achieved highest Top-1 accuracy for all models on CIFAR-10
  - Mish achieved highest Top-1 accuracy for most models on CIFAR-100

Findings

- Mish demonstrates significantly higher accuracy than ReLU and Swish
    - Trade-off: higher epoch time (due to additional computational strain)
    - Mish outperformed ReLU even with parameters optimized for ReLU
    - Smoother transitions between scalar magnitudes means smoother loss functions, which are easier to optimize

Questions

- Why is introducing non-linearity so important?