Leaky Units and Multiple Time Scales

Sargur Srihari
srihari@buffalo.edu

Topics

- Recurrent Neural Networks

1. Unfolding Computational Graphs
2. Recurrent Neural Networks
3. Bidirectional RNNs
4. Encoder-Decoder Sequence-to-Sequence Architectures
5. Deep Recurrent Networks
6. Recursive Neural Networks
7. The Challenge of Long-Term Dependencies
8. Echo-State Networks
9. Leaky Units and Other Strategies for Multiple Time Scales
10. LSTM and Other Gated RNNs
11. Optimization for Long-Term Dependencies
12. Explicit Memory

Multiple Time Scales

- One way to deal with long-term dependencies is to design a model that operates at multiple time scales
  - Some parts of the model operate at fine-grained time scales and can handle small details
  - Other parts operate at coarse time scales and transfer information from the distant past to the present more efficiently
- Strategies for building both fine and coarse time scales
  - Addition of skip connections across time
  - *Leaky units* that integrate signals with different time constants
  - Removal of some of the connections used to model fine-grained time scales

Adding skip connections through time

- One way to obtain coarse time scales is to add direct connections from variables in the distant past to variables in the present
- In an ordinary RNN, recurrent connection goes from time t to time $t+1$. Can construct RNNs with longer delays
- Gradients can vanish/explode exponentially wrt no. of time steps
- Introduce time delay of d to mitigate this problem
- Gradients diminish as a function of τ / d rather than τ
- Allows learning algorithm to capture longer dependencies
  - Not all long-term dependencies can be captured this way

Leaky units and a spectrum of time scales

- Rather than an integer skip of d time steps, the effect can be obtained smoothly by adjusting a real-valued α
- Running Average
  - Running average $\mu^{(t)}$ of some value $v^{(t)}$ is $\mu^{(t)} \leftarrow \alpha\, \mu^{(t-1)} + (1-\alpha)\, v^{(t)}$
  - Called a linear self-correction
  - When α is close to 1, running average remembers information from the past for a long time and when it is close to 0, information is rapidly discarded.
- Hidden units with linear self connections behave similar to running average. They are called *leaky units*.
- Can obtain product of derivatives close to 1 by having *linear* self-connections and a weight near 1 on those connections

Removing Connections

- Another approach to handle long-term dependencies
- Organize state of the RNN at multiple time scales
  - Information flowing more easily through long distances at the slower time scales
- It involves actively removing length one connections and replacing them with longer connections
- Skip connections add edges