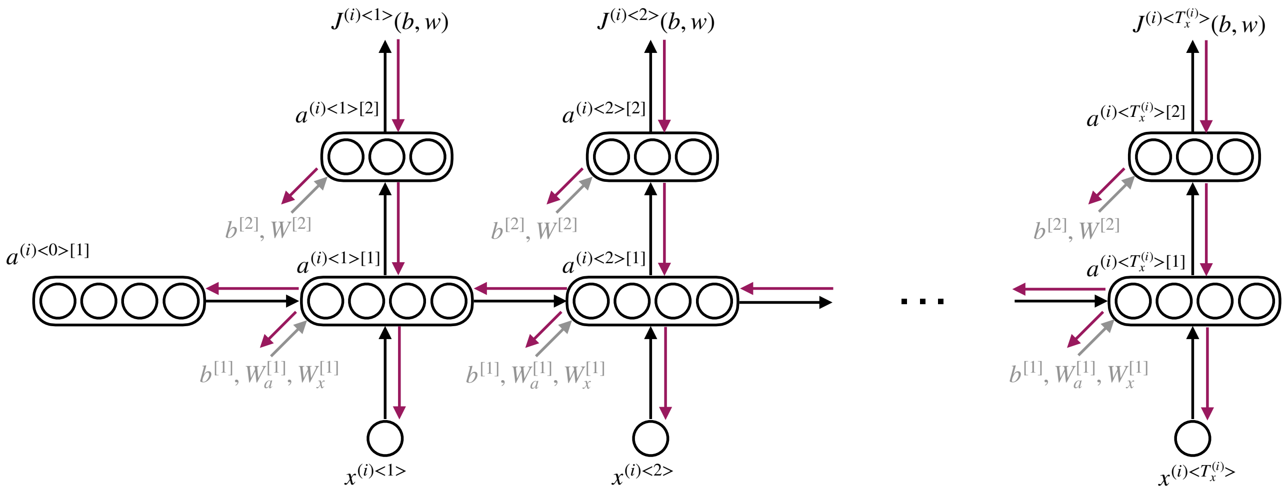
Misc. Details about RNNs

Mar. 4, 2020

Backpropagation

- In this figure, connections between layers should be illustrated as densely connected but the figure was too busy.



Forward Propagation: Start with $x^{(i)<1>}$ and $a^{(i)<0>[1]}$, work forwards in time

$$a^{(i)<t>[1]} = g^{[1]}(z^{(i)<t>[1]})$$

$$z^{(i)<t>[1]} = b^{[1]} + W_a^{[1]} a^{(i)<t-1>[1]} + W_x^{[1]} x^{(i)<t-1>}$$

Backward Propagation: Start with $J^{(i)<T_x^{(i)}>}(b,w)$, work backwards in time

For time t, $J^{(i)}(b,w) = \sum_{t=1}^{T_x^{(i)}} J^{(i)<t>}(b,w)$ depends on $a^{(i)<t>[1]}$ through $a^{(i)<t>[2]}$ (same time, next layer up) and $a^{(i)<t+1>[1]}$ (next time, same layer)

Therefore, $$\frac{\partial J^{(i)}}{\partial a^{(i)<t>[1]}} = \frac{\partial J^{(i)}}{\partial a^{(i)<t>[2]}} \frac{\partial a^{(i)<t>[2]}}{\partial a^{(i)<t>[1]}} + \frac{\partial J^{(i)}}{\partial a^{(i)<t+1>[1]}} \frac{\partial a^{(i)<t+1>[1]}}{\partial a^{(i)<t>[1]}}$$

$J^{(i)}(b,w)$ depends on $W_x^{[1]}$ through $a^{(i)<1>[1]}, \ldots, a^{(i)<T_x^{(i)}>[1]}$

Therefore, $$\frac{\partial J^{(i)}}{\partial W_x^{[1]}} = \sum_{t=1}^{T_x^{(1)}} \frac{\partial J^{(i)}}{\partial a^{(i)<t>[1]}} \frac{\partial a^{(i)<t>[1]}}{\partial W_x^{[1]}}$$ (a similar idea holds for $b^{[1]}$ and $W_a^{[1]}$)

- Note that vanishing/exploding gradients are going to be a problem:
  - We get a term in our gradient computation for every time point (e.g., for every word in our sentence).

Dependence formulation

- At time t, predictions are from $P(Y^{(i)<t>} = y^{(i)<t>} | x^{(i)<1>}, \ldots, x^{(i)<t>})$
- Depends on observed inputs up through current time.
- We calculate the joint distribution for responses across all times as follows:

$$\begin{aligned} &P(Y^{(i)<1>} = y^{(i)<1>}, \ldots, Y^{(i)<T_y^{(i)}>} = y^{(i)<T_y^{(i)}>} | x^{(i)<1>}, \ldots, x^{(i)<T_x^{(i)}>}) \\ &= P(Y^{(i)<1>} = y^{(i)<1>} | x^{(i)<1>}) \times P(Y^{(i)<2>} = y^{(i)<2>} | x^{(i)<1>}, x^{(i)<2>}) \\ &\times \cdots \times P(Y^{(i)<T_y^{(i)}>} = y^{(i)<T_y^{(i)}>} | x^{(i)<1>}, \ldots, x^{(i)<T_y^{(i)}>}) \end{aligned}$$

- This decomposition requires an assumption that $Y^{(i)<t>}$ is **conditionally independent of** $Y^{(i)<1>}, \ldots, Y^{(i)<t-1>}$ **given** $x^{(i)<1>}, \ldots, x^{(i)<T_y^{(i)}>}$.
  - Knowing the values of $Y^{(i)<1>}, \ldots, Y^{(i)<t-1>}$ wouldn't add any more information about $Y^{(i)<t>}$ than is already contained in $x^{(i)<1>}, \ldots, x^{(i)<T_y^{(i)}>}$.
- This seems restrictive, but isn't that bad.
  - We can always expand $x^{(i)<t>}$ to include the observed response at the previous time, $y^{(i)<t-1>}$, as a feature.

Multiple RNN Layers

- You can stack multiple RNN layers on top of each other.
- They could have different numbers of layers.