CS 4248
Natural Language Processing

Professor NG Hwee Tou

Department of Computer Science
School of Computing
National University of Singapore
nght@comp.nus.edu.sg

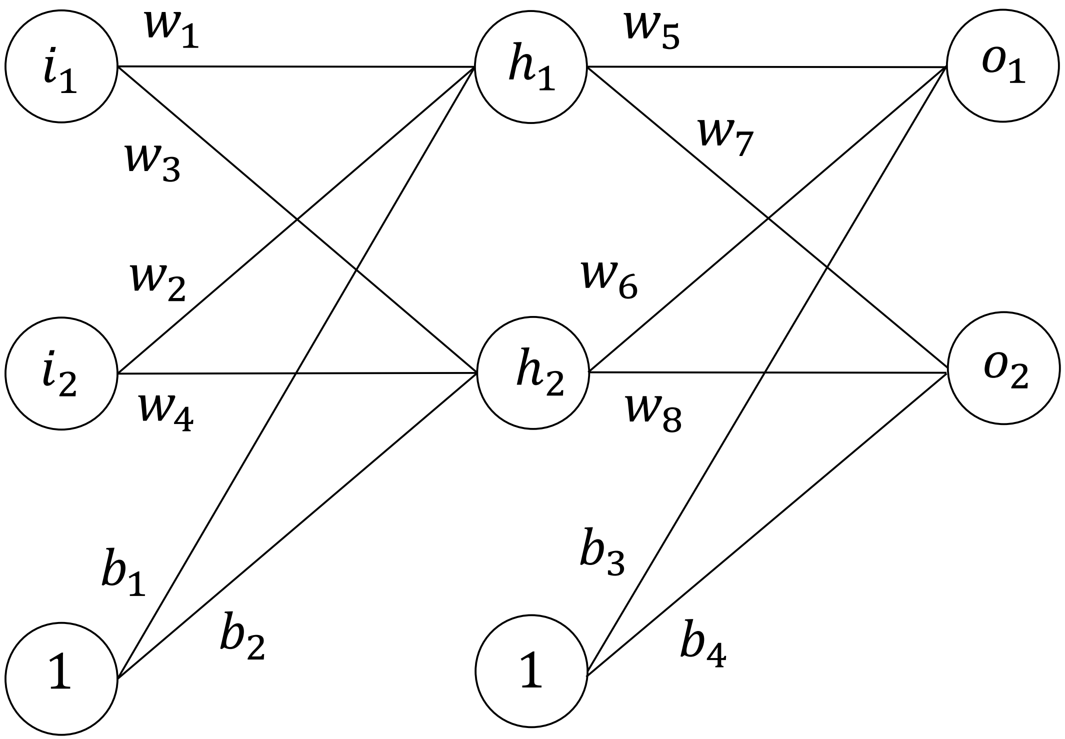
Materials

- NNM4NLP Chapter 5

Neural Network Training

- Backpropagation algorithm
- Input: a multilayer feed-forward neural network with a fixed set of units and connections
- Learns the weights for the connections
- Derivation: uses chain rule in calculus to compute the derivative of the composition of two or more functions

Backpropagation Algorithm



Training example:
Input: (i_1, i_2)
Output: (t_1, t_2)

Forward Computation

$$s_1 = w_1 i_1 + w_2 i_2 + b_1$$

$$h_1 = \frac{1}{1 + e^{-s_1}}$$

$$s_2 = w_3 i_1 + w_4 i_2 + b_2$$

$$h_2 = \frac{1}{1 + e^{-s_2}}$$

Forward Computation

$$s_3 = w_5 h_1 + w_6 h_2 + b_3$$

$$o_1 = \frac{1}{1 + e^{-s_3}}$$

$$s_4 = w_7 h_1 + w_8 h_2 + b_4$$

$$o_2 = \frac{1}{1 + e^{-s_4}}$$

$$L = \frac{1}{2}[(o_1 - t_1)^2 + (o_2 - t_2)^2]$$

Weight Update

- Gradient descent:

$$w_i \leftarrow w_i - \alpha \frac{\partial L}{\partial w_i} \qquad \alpha > 0$$

Chain Rule Theorem in Calculus

If $L = f(x_1, x_2, \dots, x_n)$ is a differentiable function of the n variables $x_1, x_2, \dots, x_n$, where each x_i is a differentiable function of the m variables $w_1, w_2, \dots, w_m$, then

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial x_1}\frac{\partial x_1}{\partial w_1} + \frac{\partial L}{\partial x_2}\frac{\partial x_2}{\partial w_1} + \cdots + \frac{\partial L}{\partial x_n}\frac{\partial x_n}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial x_1}\frac{\partial x_1}{\partial w_2} + \frac{\partial L}{\partial x_2}\frac{\partial x_2}{\partial w_2} + \cdots + \frac{\partial L}{\partial x_n}\frac{\partial x_n}{\partial w_2}$$

$\dots$

$$\frac{\partial L}{\partial w_m} = \frac{\partial L}{\partial x_1}\frac{\partial x_1}{\partial w_m} + \frac{\partial L}{\partial x_2}\frac{\partial x_2}{\partial w_m} + \cdots + \frac{\partial L}{\partial x_n}\frac{\partial x_n}{\partial w_m}$$

Backward Computation

- Base case

$$s_3 = w_5 h_1 + w_6 h_2 + b_3$$

$$o_1 = \frac{1}{1 + e^{-s_3}}$$

$$L = \frac{1}{2}[(o_1 - t_1)^2 + (o_2 - t_2)^2]$$

$$\frac{\partial L}{\partial w_5} = \frac{\partial L}{\partial s_3}\frac{\partial s_3}{\partial w_5} = \frac{\partial L}{\partial o_1}\frac{\partial o_1}{\partial s_3}\frac{\partial s_3}{\partial w_5}$$

$$= (o_1 - t_1) \times o_1(1 - o_1) \times h_1$$

Backward Computation

- Base case

$$s_3 = w_5 h_1 + w_6 h_2 + b_3$$

$$o_1 = \frac{1}{1 + e^{-s_3}}$$

$$L = \frac{1}{2}[(o_1 - t_1)^2 + (o_2 - t_2)^2]$$

$$\frac{\partial L}{\partial w_6} = \frac{\partial L}{\partial s_3}\frac{\partial s_3}{\partial w_6} = \frac{\partial L}{\partial o_1}\frac{\partial o_1}{\partial s_3}\frac{\partial s_3}{\partial w_6}$$

$$= (o_1 - t_1) \times o_1(1 - o_1) \times h_2$$

Backward Computation

- Base case

$$s_4 = w_7 h_1 + w_8 h_2 + b_4$$

$$o_2 = \frac{1}{1 + e^{-s_4}}$$

$$L = \frac{1}{2}[(o_1 - t_1)^2 + (o_2 - t_2)^2]$$

$$\frac{\partial L}{\partial w_7} = \frac{\partial L}{\partial s_4} \frac{\partial s_4}{\partial w_7} = \frac{\partial L}{\partial o_2} \frac{\partial o_2}{\partial s_4} \frac{\partial s_4}{\partial w_7}$$

$$= (o_2 - t_2) \times o_2(1 - o_2) \times h_1$$

Backward Computation

- Base case

$$s_4 = w_7 h_1 + w_8 h_2 + b_4$$

$$o_2 = \frac{1}{1 + e^{-s_4}}$$

$$L = \frac{1}{2}\left[(o_1 - t_1)^2 + (o_2 - t_2)^2\right]$$

$$\frac{\partial L}{\partial w_8} = \frac{\partial L}{\partial s_4}\frac{\partial s_4}{\partial w_8} = \frac{\partial L}{\partial o_2}\frac{\partial o_2}{\partial s_4}\frac{\partial s_4}{\partial w_8}$$

$$= (o_2 - t_2) \times o_2(1 - o_2) \times h_2$$

Backward Computation

- Recursive case

$$s_1 = w_1 i_1 + w_2 i_2 + b_1 \qquad h_1 = \frac{1}{1 + e^{-s_1}}$$

$$s_3 = w_5 h_1 + w_6 h_2 + b_3 \qquad s_4 = w_7 h_1 + w_8 h_2 + b_4$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial s_1}\frac{\partial s_1}{\partial w_1} = \frac{\partial L}{\partial h_1}\frac{\partial h_1}{\partial s_1}\frac{\partial s_1}{\partial w_1}$$

$$= \left(\frac{\partial L}{\partial s_3}\frac{\partial s_3}{\partial h_1} + \frac{\partial L}{\partial s_4}\frac{\partial s_4}{\partial h_1}\right)\frac{\partial h_1}{\partial s_1}\frac{\partial s_1}{\partial w_1}$$

$$= \left(\frac{\partial L}{\partial s_3} w_5 + \frac{\partial L}{\partial s_4} w_7\right) \times h_1(1 - h_1) \times i_1$$

$$\frac{\partial L}{\partial s_3} = (o_1 - t_1) \times o_1(1 - o_1) \qquad \frac{\partial L}{\partial s_4} = (o_2 - t_2) \times o_2(1 - o_2)$$

Backward Computation

- Recursive case

$$s_1 = w_1 i_1 + w_2 i_2 + b_1$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial s_1}\frac{\partial s_1}{\partial w_2}$$
$$= \left(\frac{\partial L}{\partial s_3} w_5 + \frac{\partial L}{\partial s_4} w_7\right) \times h_1(1 - h_1) \times i_2$$

Backward Computation

- Recursive case

$$s_2 = w_3 i_1 + w_4 i_2 + b_2 \qquad h_2 = \frac{1}{1 + e^{-s_2}}$$

$$s_3 = w_5 h_1 + w_6 h_2 + b_3 \qquad s_4 = w_7 h_1 + w_8 h_2 + b_4$$

$$\frac{\partial L}{\partial w_3} = \frac{\partial L}{\partial s_2}\frac{\partial s_2}{\partial w_3} = \frac{\partial L}{\partial h_2}\frac{\partial h_2}{\partial s_2}\frac{\partial s_2}{\partial w_3}$$

$$= \left(\frac{\partial L}{\partial s_3}\frac{\partial s_3}{\partial h_2} + \frac{\partial L}{\partial s_4}\frac{\partial s_4}{\partial h_2}\right)\frac{\partial h_2}{\partial s_2}\frac{\partial s_2}{\partial w_3}$$

$$= \left(\frac{\partial L}{\partial s_3} w_6 + \frac{\partial L}{\partial s_4} w_8\right) \times h_2(1 - h_2) \times i_1$$

$$\frac{\partial L}{\partial s_3} = (o_1 - t_1) \times o_1(1 - o_1) \qquad \frac{\partial L}{\partial s_4} = (o_2 - t_2) \times o_2(1 - o_2)$$

Backward Computation

- Recursive case

$$s_2 = w_3 i_1 + w_4 i_2 + b_2$$

$$\frac{\partial L}{\partial w_4} = \frac{\partial L}{\partial s_2}\frac{\partial s_2}{\partial w_4}$$

$$= (\frac{\partial L}{\partial s_3} w_6 + \frac{\partial L}{\partial s_4} w_8) \times h_2(1 - h_2) \times i_2$$

Example

$w_1 = 0.15 \quad w_2 = 0.25 \quad w_3 = 0.20 \quad w_4 = 0.30$

$w_5 = 0.10 \quad w_6 = 0.35 \quad w_7 = 0.05 \quad w_8 = -0.20$

$b_1 = -0.15 \quad b_2 = -0.10 \quad b_3 = -0.50 \quad b_4 = -0.20$

$i_1 = 0.20 \quad i_2 = 0.50 \quad t_1 = 1.00 \quad t_2 = 0.00$

$\alpha = 0.50$

Example

$$s_1 = w_1 i_1 + w_2 i_2 + b_1$$

$$= 0.15 \times 0.20 + 0.25 \times 0.50 - 0.15 = 0.005$$

$$h_1 = \frac{1}{1 + e^{-s_1}} = \frac{1}{1 + e^{-0.005}} = 0.5013$$

$$s_2 = w_3 i_1 + w_4 i_2 + b_2$$

$$= 0.20 \times 0.20 + 0.30 \times 0.50 - 0.10 = 0.09$$

$$h_2 = \frac{1}{1 + e^{-s_2}} = \frac{1}{1 + e^{-0.09}} = 0.5225$$

Example

$$s_3 = w_5 h_1 + w_6 h_2 + b_3$$

$$= 0.10 \times 0.5013 + 0.35 \times 0.5225 - 0.50 = -0.2670$$

$$o_1 = \frac{1}{1 + e^{-s_3}} = \frac{1}{1 + e^{0.2670}} = 0.4336$$

$$s_4 = w_7 h_1 + w_8 h_2 + b_4$$

$$= 0.05 \times 0.5013 - 0.20 \times 0.5225 - 0.20 = -0.2794$$

$$o_2 = \frac{1}{1 + e^{-s_4}} = \frac{1}{1 + e^{0.2794}} = 0.4306$$

Example

$$L = \frac{1}{2}\left[(o_1 - t_1)^2 + (o_2 - t_2)^2\right]$$

$$= \frac{1}{2}\left[(0.4336 - 1.00)^2 + (0.4306 - 0.00)^2\right] = 0.2531$$

Example

$$\frac{\partial L}{\partial s_3} = (o_1 - t_1) \times o_1(1 - o_1)$$

$$= (0.4336 - 1.00) \times 0.4336 \times (1 - 0.4336) = -0.1391$$

$$\frac{\partial L}{\partial s_4} = (o_2 - t_2) \times o_2(1 - o_2)$$

$$= (0.4306 - 0.00) \times 0.4306 \times (1 - 0.4306) = 0.1056$$

Example

$$\frac{\partial L}{\partial w_5} = \frac{\partial L}{\partial s_3} \times h_1 = -0.1391 \times 0.5013 = -0.06972$$

$$w_5' = w_5 - \alpha \frac{\partial L}{\partial w_5} = 0.10 - 0.5 \times (-0.06972) = 0.1349$$

$$\frac{\partial L}{\partial w_7} = \frac{\partial L}{\partial s_4} \times h_1 = 0.1056 \times 0.5013 = 0.05292$$

$$w_7' = w_7 - \alpha \frac{\partial L}{\partial w_7} = 0.05 - 0.5 \times 0.05292 = 0.02354$$

Example

$$\frac{\partial L}{\partial w_1} = (\frac{\partial L}{\partial s_3} w_5 + \frac{\partial L}{\partial s_4} w_7) \times h_1(1 - h_1) \times i_1$$

$$= (-0.1391 \times 0.10 + 0.1056 \times 0.05)$$

$$\times 0.5013 \times (1 - 0.5013) \times 0.20 = -0.0004315$$

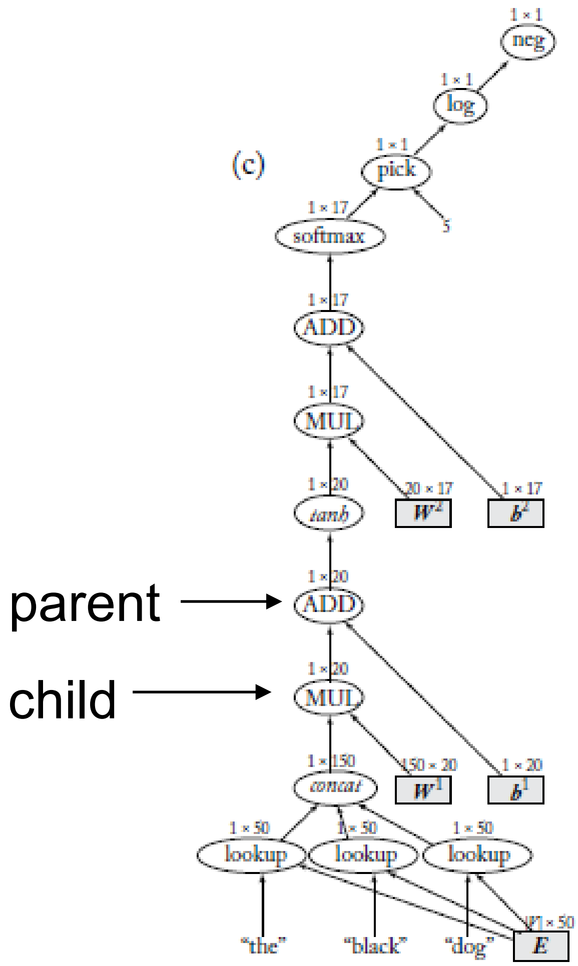
$$w_1' = w_1 - \alpha \frac{\partial L}{\partial w_1} = 0.15 - 0.5 \times (-0.0004315) = 0.1502$$

Neural Network Training

- Computation graph
  - A representation of an arbitrary mathematical computation as a graph
  - A directed acyclic graph (DAG)
    - Nodes: mathematical operations or bound variables
    - Edges: flow of intermediary values between nodes
- A neural network is essentially a mathematical expression, and can be represented as a computation graph

Neural Network Training

- Computation graph abstraction



Software

- Software packages that implement the computation graph model:
  - PyTorch (Facebook)
  - TensorFlow (Google)
  - Keras (provide a higher level interface on top of TensorFlow)

Neural Network Training

```
Define network parameters.
for iteration = 1 to T do
    for Training example x_i , y_i in dataset do
        loss_node ← build_computation_graph(x_i , y_i , parameters)
        loss_node.forward()
        gradients ← loss_node().backward()
        parameters ← update_parameters(parameters, gradients)
return parameters.
```

Convergence

- Only guaranteed to converge toward some local minimum and not necessarily the global minimum
- But in practice, it is a highly effective function approximation method

Practicalities

- Restarts
  - Different random initializations are likely to result in different final solutions and different accuracies
  - Random restarts: Run the training process multiple times, each with a different random initialization, and choose the best one on the development set

Practicalities

- Ensemble of multiple models
  - Build a different model using a different set of random initializations
  - Combine the multiple models in prediction
    - Take the majority vote of the different models
    - Average the output vectors of the different models
  - Using an ensemble of models often increases the prediction accuracy

Practicalities

- Learning rate
  - Too large: not converging
  - Too small: taking too long to converge
  - Experiment with a range of learning rates: 0.1, 0.01, 0.001, etc.