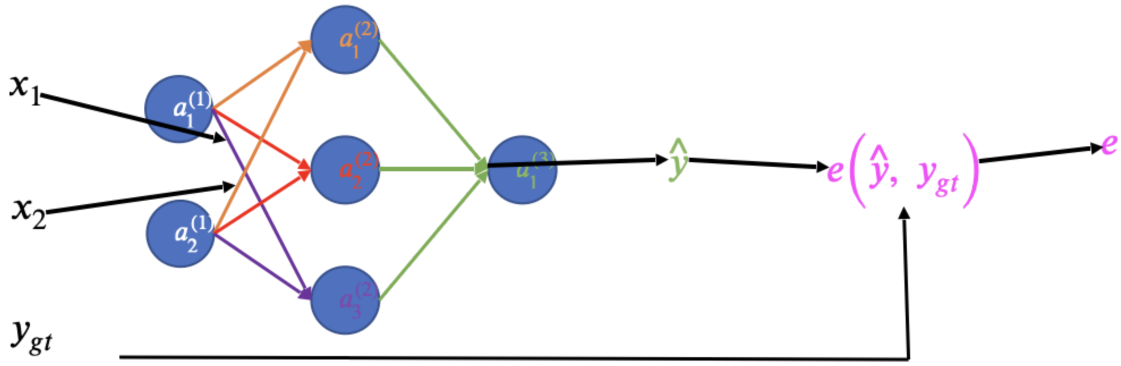
CAS CS 440
Lec 26

Supervised Learning VIII – Neural Networks (cont.)

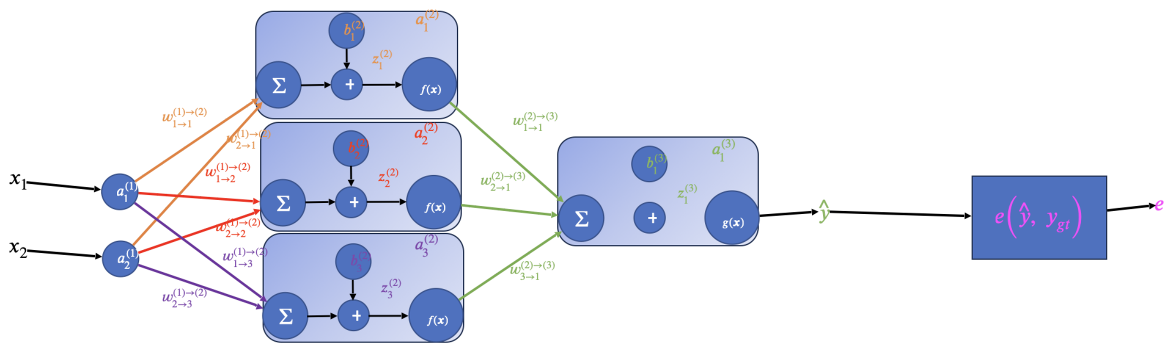
1. Example

a.



b. $$f(x) = \sigma(x) = \frac{1}{1 + e^{-x}} \qquad g(x) = I(x) = x \qquad e\left(\hat{y},\ y_{gt}\right) = \left(y_{gt} - \hat{y}\right)^2$$

c.



Forward equations:

$$z_1^{(2)} = w_{1\to1}^{(1)\to(2)} a_1^{(1)} + w_{2\to1}^{(1)\to(2)} a_2^{(1)} + b_1^{(2)} \qquad a_1^{(2)} = f\left(z_1^{(2)}\right)$$

$$z_2^{(2)} = w_{1\to2}^{(1)\to(2)} a_1^{(1)} + w_{2\to2}^{(1)\to(2)} a_2^{(1)} + b_2^{(2)} \qquad a_2^{(2)} = f\left(z_2^{(2)}\right)$$

$$z_3^{(2)} = w_{1\to3}^{(1)\to(2)} a_1^{(1)} + w_{2\to3}^{(1)\to(2)} a_2^{(1)} + b_3^{(2)} \qquad a_3^{(2)} = f\left(z_3^{(2)}\right)$$

d. $$z_1^{(3)} = w_{1\to1}^{(2)\to(3)} a_1^{(2)} + w_{2\to1}^{(2)\to(3)} a_2^{(2)} + w_{3\to1}^{(2)\to(3)} a_2^{(2)} + b_1^{(3)} \qquad a_1^{(3)} = f\left(z_1^{(3)}\right)$$

Backward equations:

$$\frac{\partial e}{\partial w^{(2)\to(3)}_{2\to1}} = \frac{\partial e}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z_1^{(3)}}\frac{\partial z_1^{(3)}}{\partial w^{(2)\to(3)}_{2\to1}} = -\hat{y}\left(y_{gt}-\hat{y}\right)^1\ a_2^{(2)}$$

$\bullet\bullet\bullet$

$$\frac{\partial e}{\partial w^{(1)\to(2)}_{1\to1}} = \frac{\partial e}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z_1^{(3)}}\frac{\partial z_1^{(3)}}{\partial a_1^{(2)}}\frac{\partial a_1^{(2)}}{\partial z_1^{(2)}}\frac{\partial z_1^{(2)}}{\partial w^{(1)\to(2)}_{1\to1}} = -\hat{y}\left(y_{gt}-\hat{y}\right)^1\ w^{(2)\to(3)}_{1\to1}\ \sigma'\left(z_1^{(2)}\right)$$

$\bullet\bullet\bullet$

$$\frac{\partial e}{\partial w^{(1)\to(2)}_{1\to2}} = \frac{\partial e}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z_1^{(3)}}\frac{\partial z_1^{(3)}}{\partial a_2^{(2)}}\frac{\partial a_2^{(2)}}{\partial z_2^{(2)}}\frac{\partial z_2^{(2)}}{\partial w^{(1)\to(2)}_{1\to2}} = -\hat{y}\left(y_{gt}-\hat{y}\right)^1\ w^{(2)\to(3)}_{2\to1}\ \sigma'\left(z_2^{(2)}\right)$$

$\bullet\bullet\bullet$

$$\frac{\partial e}{\partial w^{(1)\to(2)}_{2\to3}} = \frac{\partial e}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z_1^{(3)}}\frac{\partial z_1^{(3)}}{\partial a_3^{(2)}}\frac{\partial a_3^{(2)}}{\partial z_3^{(2)}}\frac{\partial z_3^{(2)}}{\partial w^{(1)\to(2)}_{2\to3}} = -\hat{y}\left(y_{gt}-\hat{y}\right)^1\ w^{(2)\to(3)}_{2\to1}\ \sigma'\left(z_3^{(2)}\right)$$

e. $\bullet\bullet\bullet$

$$\frac{\partial e}{\partial b_1^{(3)}} = \frac{\partial e}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z_1^{(3)}}\frac{\partial z_1^{(3)}}{\partial b_1^{(3)}} = -\hat{y}\left(y_{gt}-\hat{y}\right)^1 1$$

$$\frac{\partial e}{\partial b_1^{(2)}} = \frac{\partial e}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z_1^{(3)}}\frac{\partial z_1^{(3)}}{\partial a_1^{(2)}}\frac{\partial a_1^{(2)}}{\partial z_1^{(2)}}\frac{\partial z_1^{(2)}}{\partial b_1^{(2)}} = -\hat{y}\left(y_{gt}-\hat{y}\right)^1\ w^{(2)\to(3)}_{1\to1}\ \sigma'\left(z_1^{(2)}\right)$$

Backward equations:

$$\frac{\partial e}{\partial w^{(2)\to(3)}_{2\to1}} = \frac{\partial e}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z_1^{(3)}}\frac{\partial z_1^{(3)}}{\partial w^{(2)\to(3)}_{2\to1}} \quad \bullet\bullet\bullet \qquad \frac{\partial e}{\partial b_1^{(3)}} = \frac{\partial e}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z_1^{(3)}}\frac{\partial z_1^{(3)}}{\partial b_1^{(3)}}$$

Shared w/ other $\frac{\partial e}{\partial w^{(1)\to(2)}_{i\to j}}$

$$\frac{\partial e}{\partial w^{(1)\to(2)}_{1\to1}} = \frac{\partial e}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z_1^{(3)}}\frac{\partial z_1^{(3)}}{\partial a_1^{(2)}}\frac{\partial a_1^{(2)}}{\partial z_1^{(2)}}\frac{\partial z_1^{(2)}}{\partial w^{(1)\to(2)}_{1\to1}} \quad \bullet\bullet\bullet \qquad \frac{\partial e}{\partial b_1^{(2)}} = \frac{\partial e}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z_1^{(3)}}\frac{\partial z_1^{(3)}}{\partial a_1^{(2)}}\frac{\partial a_1^{(2)}}{\partial z_1^{(2)}}\frac{\partial z_1^{(2)}}{\partial b_1^{(2)}}$$

Shared w/ other $\frac{\partial e}{\partial w^{(1)\to(2)}_{i\to j}}$

$$\frac{\partial e}{\partial w^{(1)\to(2)}_{1\to2}} = \frac{\partial e}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z_1^{(3)}}\frac{\partial z_1^{(3)}}{\partial a_2^{(2)}}\frac{\partial a_2^{(2)}}{\partial z_2^{(2)}}\frac{\partial z_2^{(2)}}{\partial w^{(1)\to(2)}_{1\to2}} \quad \bullet\bullet\bullet \qquad \frac{\partial e}{\partial b_2^{(2)}} = \frac{\partial e}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z_1^{(3)}}\frac{\partial z_1^{(3)}}{\partial a_2^{(2)}}\frac{\partial a_2^{(2)}}{\partial z_2^{(2)}}\frac{\partial z_2^{(2)}}{\partial b_2^{(2)}}$$

$$\frac{\partial e}{\partial w^{(1)\to(2)}_{2\to3}} = \frac{\partial e}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z_1^{(3)}}\frac{\partial z_1^{(3)}}{\partial a_3^{(2)}}\frac{\partial a_3^{(2)}}{\partial z_3^{(2)}}\frac{\partial z_3^{(2)}}{\partial w^{(1)\to(2)}_{2\to3}} \quad \bullet\bullet\bullet \qquad \frac{\partial e}{\partial b_3^{(2)}} = \frac{\partial e}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z_1^{(3)}}\frac{\partial z_1^{(3)}}{\partial a_3^{(2)}}\frac{\partial a_3^{(2)}}{\partial z_3^{(2)}}\frac{\partial z_3^{(2)}}{\partial b_3^{(2)}}$$

Only compute once! Shared w/ other $\frac{\partial e}{\partial w^{(1)\to(2)}_{i\to j}}$ Only compute once!

f.

i. Cache the value and compute it only once and store them

$$\frac{\partial e}{\partial z_1^{(3)}} = \frac{\partial e}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_1^{(3)}} \qquad = -\hat{y}\left(y_{gt} - \hat{y}\right) g'\left(z_1^{(3)}\right)$$

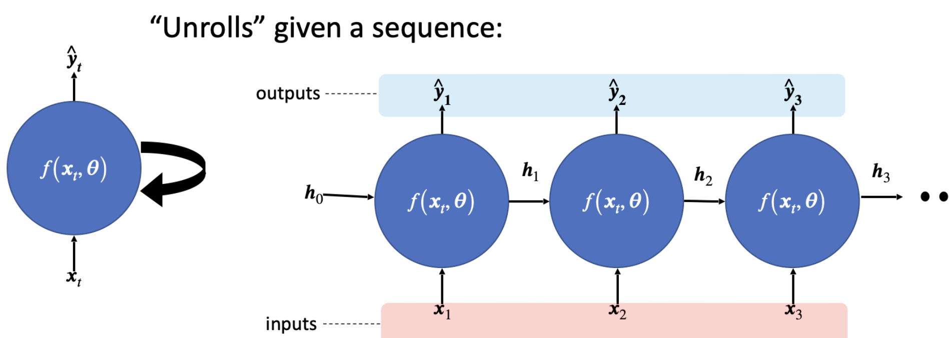
$$\frac{\partial e}{\partial z_1^{(2)}} = \frac{\partial e}{\partial z_1^{(3)}} \frac{\partial z_1^{(3)}}{\partial a_1^{(2)}} \frac{\partial a_1^{(2)}}{\partial z_1^{(2)}} \qquad = \frac{\partial e}{\partial z_1^{(3)}} w_{1\to 1}^{(2)\to(3)} \sigma'\left(z_1^{(2)}\right)$$

$$\frac{\partial e}{\partial z_2^{(2)}} = \frac{\partial e}{\partial z_1^{(3)}} \frac{\partial z_1^{(3)}}{\partial a_2^{(2)}} \frac{\partial a_2^{(2)}}{\partial z_2^{(2)}} \qquad = \frac{\partial e}{\partial z_1^{(3)}} w_{2\to 1}^{(2)\to(3)} \sigma'\left(z_2^{(2)}\right)$$

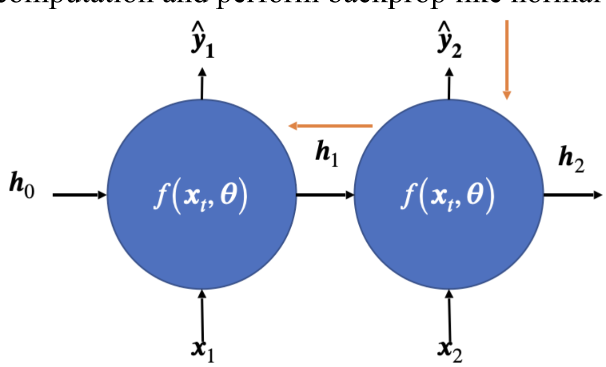
g. $$\frac{\partial e}{\partial z_3^{(2)}} = \frac{\partial e}{\partial z_1^{(3)}} \frac{\partial z_1^{(3)}}{\partial a_3^{(2)}} \frac{\partial a_3^{(2)}}{\partial z_3^{(2)}} \qquad = \frac{\partial e}{\partial z_1^{(3)}} w_{3\to 1}^{(2)\to(3)} \sigma'\left(z_3^{(2)}\right)$$

   i. Cached values

2. Backpropagation Through Time (BPTT)
   a. What about sequences? Current NNs only deal with grids/vector input
   b. Recurrent NNs (RNNs): derived from State Machines
   c. “Unrolls” given a sequence

      i. 

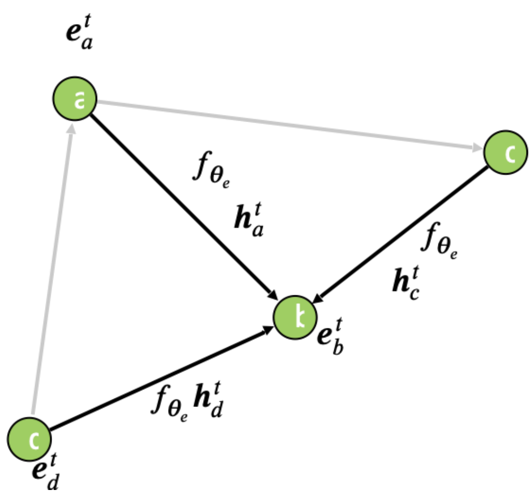
3. Vanilla RNN with BPTT
   a. Using computation and perform backprop like normal



      i. $$\begin{aligned} \boldsymbol{H}_t &= f\left(\boldsymbol{X}_t \boldsymbol{U} + \boldsymbol{H}_{t-1} \boldsymbol{W} + \boldsymbol{b}_s\right) \\ \boldsymbol{Y}_t &= \boldsymbol{H}_t \boldsymbol{V} + \boldsymbol{b}_o \end{aligned}$$

1. Input is in batch form

ii. $$\nabla_\theta L = \begin{pmatrix} \frac{\partial L}{\partial \boldsymbol{b}_o} \\ \frac{\partial L}{\partial \boldsymbol{b}_s} \\ \frac{\partial L}{\partial \boldsymbol{U}} \\ \frac{\partial L}{\partial \boldsymbol{W}} \\ \frac{\partial L}{\partial \boldsymbol{V}} \end{pmatrix} = \begin{pmatrix} \frac{\partial L}{\partial \boldsymbol{b}_o} \mid x_2, y_2 \\ \frac{\partial L}{\partial \boldsymbol{b}_s} \mid x_2, y_2 \\ \frac{\partial L}{\partial \boldsymbol{U}} \mid x_2, y_2 \\ \frac{\partial L}{\partial \boldsymbol{W}} \mid x_2, y_2 \\ \frac{\partial L}{\partial \boldsymbol{V}} \mid x_2, y_2 \end{pmatrix} + \begin{pmatrix} \frac{\partial L}{\partial \boldsymbol{b}_o} \mid x_1, y_1 \\ \frac{\partial L}{\partial \boldsymbol{b}_s} \mid x_1, y_1 \\ \frac{\partial L}{\partial \boldsymbol{U}} \mid x_1, y_1 \\ \frac{\partial L}{\partial \boldsymbol{W}} \mid x_1, y_1 \\ \frac{\partial L}{\partial \boldsymbol{V}} \mid x_1, y_1 \end{pmatrix}$$

4. RNN Variants
    a. Different “cell types”
        i. LSTM
        ii. GRU
        iii. LRU
        iv. …
    b. Only difference is in how internal state is computed/maintained
        i. Big difference, add specific areas for “forgetting” and “remembering”
5. Graph Structure?
    a. 

    b. What you need to define:

An edge function f_{θ_e}

An aggregator function γ

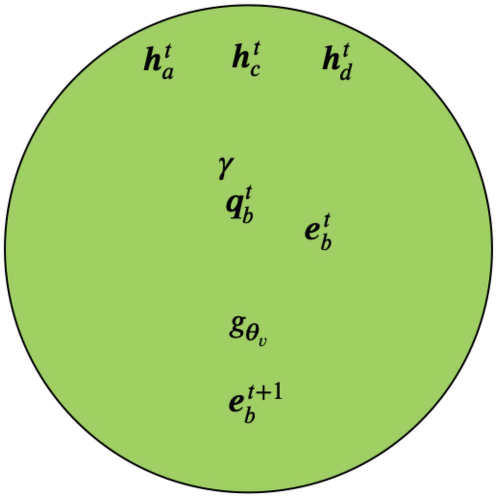
A vertex function g_{θ_v}

Computational Process:

1) $\boldsymbol{h}_i^t = f_{\theta_e}(\boldsymbol{e}_i^t),\ i \in N(b)$

2) $\boldsymbol{q}_b^t = \gamma\left(\{\boldsymbol{h}_i^t\}_{i \in N(b)}\right)$

3) $\boldsymbol{e}_b^{t+1} = g_{\theta_v}(\boldsymbol{e}_b^t,\ \boldsymbol{q}_b^t)$



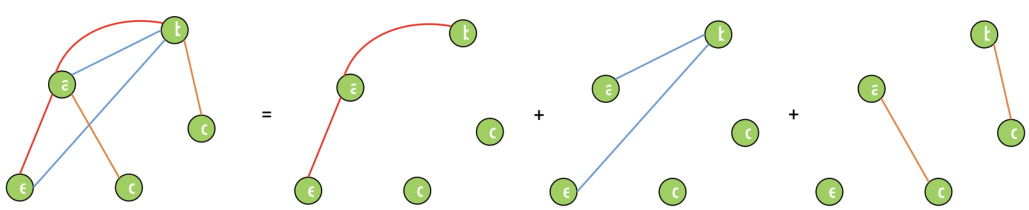
c.

6. Easy to Vectorize
    a. $\gamma = \sum$ (matrix mult. w/ adjacency matrix)
    b. f_{θ_e} = dense NN
    c. g_{θ_v} = some (element-wise independent) activation function ‘g’
    d. next layer

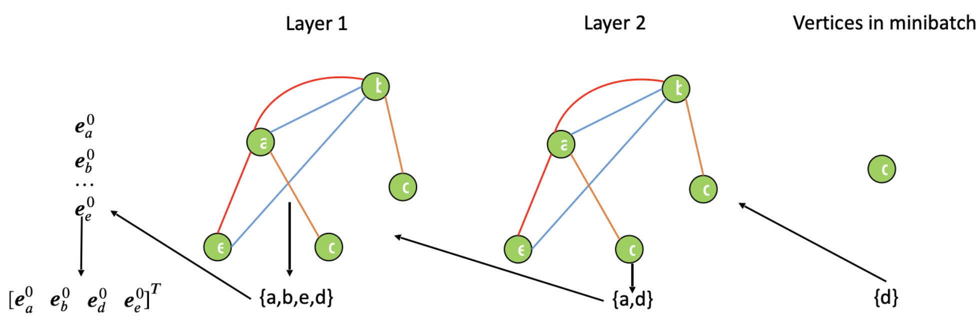
$$E^{t+1} = g(A^t E^t \overbrace{W^t}^{f_{\theta_e}})$$

vectors for all vertices at current layer

adjacency matrix for current layer

7. Multi-view Graphs?
    a. Edges all have one “type” (represent the same type of link)
    b. Not true for many problems
    c. Can we make GNNs work for multi-view graphs?
    d. 
    e. $$E^{t+1} = g\left(\frac{1}{3}A^t E^t W^t + \frac{1}{3}A^t E^t W^t + \frac{1}{3}A^t E^t W^t\right)$$
8. Can we use GNNs to model the brain?

a. Built in topology
b. Encode lifecycle rules
    i. Start off with lots of connections
    ii. Prune connections over time
c. Natural architecture for wiring costs, energy constraints, etc.
d. 

9. Make GNNs Computable
    a. $\boldsymbol{E}^{t+1} = g\left(\boldsymbol{A}^t \boldsymbol{E}^t \boldsymbol{W}^t\right)$ takes a lot of memory (and time)
    b. Problem: too big for (most) GPUs….doesn't scale
        i. Limitations in sparse support
    c. Can we make it?
    d. 

10. Changing the Graph Structure: Vertex Inflation
    a. The brain changes its structure over its lifetime
    b. GNNs reuse the same graph for each layer
    c. Idea: change the graph structure as a function of layer
        i. Vertex inflation: shortcut "long paths" in the graph
        ii. Performance?
    d.