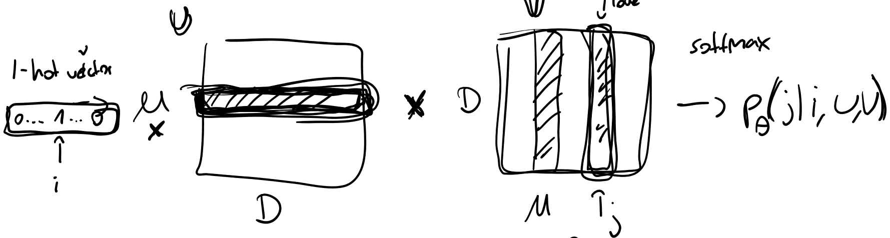
Wednesday, 10 May 2023 13:17

bat, car

bat → i

$h_\theta(x^{(t)}, x^{(t-1)}, x^{(t+1)} \dots)$ → RNN

Problem 1: Word2vec defines a mapping from a single word to a single fixed vector. Explain and provide an example why this will not be expressive enough regarding homographs (i.e., words with the same spelling but having more than one meaning). Propose an alternative solution.



Training on pairs:

I love MLCS

$S \{ (I, love), (love, I), (love, MLCS), (MLCS, love), (I, MLCS), (MLCS, I) \dots$

objective:

$\max \prod_{i,j \in S} p(j \mid i, \theta)$

$\theta = \{U, V\}$

$h_\theta(x^{(t)}, x^{(t-1)}, x^{(t-2)} \dots x^{(1)}) = p_\theta(x^{(t+1)} \mid \dots, \theta)$

Problem 2: Given a previous hidden state $\boldsymbol{h}^{(t-1)} \in \mathbb{R}^D$ and a current input $\boldsymbol{x}^{(t)} \in \mathbb{R}^N$, the recurrent neural network equations to update the hidden state and produce the output are:

$$\boldsymbol{z}^{(t)} = \boldsymbol{W}\boldsymbol{h}^{(t-1)} + \boldsymbol{U}\boldsymbol{x}^{(t)}$$
$$\boldsymbol{h}^{(t)} = \tanh\left(\boldsymbol{z}^{(t)}\right)$$
$$\boldsymbol{o}^{(t)} = \boldsymbol{V}\boldsymbol{h}^{(t)}$$
$$\hat{\boldsymbol{y}}^{(t)} = \operatorname{softmax}\left(\boldsymbol{o}^{(t)}\right)$$

where parameters $\boldsymbol{W} \in \mathbb{R}^{D\times D}$, $\boldsymbol{U} \in \mathbb{R}^{D\times N}$ and $\boldsymbol{V} \in \mathbb{R}^{M\times D}$ are shared at every step.

To train an RNN we need gradients of loss w.r.t. the parameters: $\partial L/\partial \boldsymbol{W}$, $\partial L/\partial \boldsymbol{U}$ and $\partial L/\partial \boldsymbol{V}$. Your task is to arrive at the equations given on slide 17 in the lecture.

Use the fact that $\boxed{\partial L/\partial \boldsymbol{o}^{(t)}} = \hat{\boldsymbol{y}}^{(t)} - \boldsymbol{y}^{(t)}$, where $\boldsymbol{y}^{(t)}$ is the true output.

Hint: Since parameters are shared, the total gradient is the sum of the contributions over all the steps. Because of that, it might be easier to introduce copies of parameters, e.g. $\boldsymbol{W}^{(t)}$ – a copy of $\boldsymbol{W}$ at step t, calculate $\partial L/\partial \boldsymbol{W}^{(t)}$ and sum over all t.

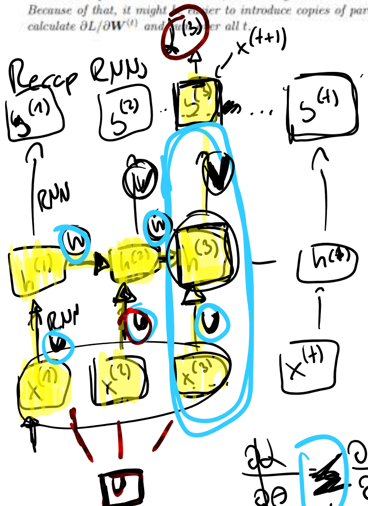
$$\frac{\partial \mathcal{L}}{\partial \Theta} = \frac{\partial \mathcal{L}}{\partial o^{(t)}} \cdot \frac{\partial o^{(t)}}{\partial \Theta}$$

$$z^{(t)} = W h^{(t-1)} + (\text{const in } W)$$

$$\boxed{\delta^{(t)} = \hat{s}^{(t)} - s^{(t)}} = \frac{\partial \mathcal{L}}{\partial o^{(t)}}$$

=

Recap RNNs



$z^{(t)} = U x^{(t)} + W h^{(t-1)}$

$h^{(t)} = \tanh(z^{(t)})$

$o^{(t)} = V h^{(t)}$

$s^{(t)} = \text{softmax}(o^{(t)})$

$\theta - \{V, U, W\}$

$\mathcal{L}(x^{(1)} \dots x^{(t)}, s^{(t)}, \theta) = \dots$ $\frac{\partial \mathcal{L}}{\partial \theta}$

$\frac{\partial \mathcal{L}}{\partial W}$ $\frac{\partial \mathcal{L}}{\partial V}$ $\frac{\partial \mathcal{L}}{\partial U}$

objective $\prod_t p(s^{(t)} \mid x^{(1)} \dots x^{(t)})$

$\to \sum_t \log p(s^{(t)} \mid x^{(1)} \dots$

$$\frac{\partial \mathcal{L}}{\partial \theta} = \sum_t \frac{\partial \mathcal{L}^{(t)}}{\partial \theta} = \boxed{\sum_t \sum_{t'} \frac{\partial \mathcal{L}^{(t)}}{\partial \theta^{(t')}}}$$

$$\frac{\partial \mathcal{L}^{(t)}}{\partial \theta} = \frac{\partial \mathcal{L}^{(t)}}{\partial U} = \sum_{\text{paths}} \frac{\partial \mathcal{L}^{(t)}}{\partial U(\text{path})} = \sum_{t'} \frac{\partial \mathcal{L}^{(t)}}{\partial U^{(t')}} = \frac{\partial \mathcal{L}^{(t)}}{\partial U^{(1)}} \; \cancel{\frac{\partial \mathcal{L}^{(4)}}{\partial U^{(3)}}} = 0$$

(1) V: $\frac{\partial \mathcal{L}}{\partial o^{(t)}} \boxed{\frac{\partial o^{(t)}}{\partial V}}$ — $\mathbb{R}^n$ — $\mathbb{R}^{n \times m}$

before: $\frac{\partial o^{(t)}}{\partial U}$

$= \frac{\partial}{\partial U} (V h^{(t)})$

x

B

equivalent:

Recap: $g: \mathbb{R}^m \to \mathbb{R}^n$ $g(x)$

$$\frac{\partial g}{\partial x} = J_g^T = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_m} \\ \vdots & \ddots & \\ \frac{\partial g_n}{\partial x_1} & \cdots & \frac{\partial g_n}{\partial x_m} \end{pmatrix}^T$$

$f: \mathbb{R}^{m \times o} \to \mathbb{R}^n$ $f(X)$

$$\left[\frac{\partial f}{\partial X}\right]_{ijh} = \frac{\partial f_i}{\partial X_{jh}}$$

$$\frac{\partial g}{\partial x} = J_g^T = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_m} \\ \vdots & \ddots & \\ \frac{\partial g_n}{\partial x_1} & \cdots & \frac{\partial g_n}{\partial x_m} \end{pmatrix} \qquad \left[\frac{\partial f}{\partial X}\right]_{jh} = \frac{\partial f_i}{\partial X_{jh}}$$

equivalent: $\frac{\partial}{\partial B} x$

$$\frac{\partial f_i}{\partial X_{jh}} \quad X_{jk}: \boxed{\frac{\partial f}{\partial X_{jh}} = \sum_i \frac{\partial f_i}{\partial X_{jh}}} \qquad \sum_i \boxed{\frac{\partial o^{(t)}_i}{\partial V_{jk}}}$$

$$\delta\left(\ (t) \right) \quad \frac{\partial L}{\partial a}$$

If smth of the form

$$\frac{\partial L}{\partial \vec{a}} \ \frac{\partial \vec{a}}{\partial B} = \frac{\partial L}{\partial \vec{a}} \ \vec{x}^T \qquad \boxed{\vec{a} = B\vec{x}}$$

$$\frac{\partial o^{(t)}_k}{\partial V_{ij}} = \frac{\partial}{\partial V_{ij}} \left(V h^{(t)}\right)_k = \frac{\partial}{\partial V_{ij}} \left[\sum_l V_{kl} h^{(t)}_l \right]_k = \sum_l \frac{\partial}{\partial V_{ij}} V_{kl} h^{(t)}_l$$

$$= \sum_l h^{(t)}_l \frac{\partial}{\partial V_{ij}} V_{kl} = \sum_{l \neq j} h^{(t)}_l \underbrace{\frac{\partial}{\partial V_{ij}} V_{kl}}_{=0} + h^{(t)}_j \frac{\partial}{\partial V_{ij}} V_{kj} = h^{(t)}_j \frac{\partial}{\partial V_{ij}} V_{kj}$$

$$\begin{cases} 0 & \text{if } i \neq k \text{ or } j \neq l \\ 1 & \text{if } i = k \text{ and } j = l \end{cases}$$

$$\frac{\partial o^{(t)}}{\partial V_{ij}} = \sum_k \frac{\partial o^{(t)}_k}{\partial V_{ij}} = \sum_k h^{(t)}_j \boxed{\frac{\partial}{\partial V_{ij}} V_{kj}}$$

$$= \sum_{k \neq i} h^{(t)}_j \underbrace{\frac{\partial}{\partial V_{ij}} V_{kl}}_{=0} + h^{(t)}_j \underbrace{\frac{\partial}{\partial V_{ij}} V_{ij}}_{=1} = \boxed{h^{(t)}_j} = \boxed{\frac{\partial o^{(t)}}{\partial V_{ij}}}$$

$=1$ if $i=k$

$$\boxed{\frac{\partial L}{\partial V_{ij}}} = \underbrace{\frac{\partial L}{\partial o^{(t)}}}_{\mathbb{R}^n} \underbrace{\frac{\partial o^{(t)}}{\partial V_{ij}}}_{\mathbb{R}^m} = \sum_k \frac{\partial L}{\partial o^{(t)}_k} \frac{\partial o^{(t)}_k}{\partial V_{ij}} = \sum_k \delta^{(t)}_k h^{(t)}_j \boxed{\frac{\partial}{\partial V_{ij}} V_{kj}}$$

$$= \boxed{\delta^{(t)}_i h^{(t)}_j}$$

$$\frac{\partial L^{(t)}}{\partial V} = \delta^{(t)} \left(h^{(t)}\right)^T$$

$$\frac{\partial L}{\partial V} = \sum_t \frac{\partial L^{(t)}}{\partial V} = \sum_t \delta^{(t)} \left(h^{(t)}\right)^T$$

δ'

$$\delta = \delta^{(t)} = \frac{\partial L}{\partial o^{(t)}} \qquad z^{(t)} = W h^{(t-1)}$$

$$\frac{\partial L}{\partial W} = \sum_t \frac{\partial L}{\partial W^{(t)}} = \sum_t \frac{\partial L}{\partial h^{(t)}_i} \frac{\partial h^{(t)}_i}{\partial z^{(t)}_j} \boxed{\frac{\partial z^{(t)}_j}{\partial W_{kl}}}$$

i) $$\frac{\partial z^{(t)}}{\partial W} = \frac{\partial}{\partial W} \left(W h^{(t-1)} + \cancel{\ldots} \right) = \frac{\partial}{\partial W} \left(W h^{(t-1)} \right)$$

but we know what "effect this has"

$$= \sum_t \boxed{\frac{\partial L}{\partial h^{(t)}} \boxed{\frac{\partial h^{(t)}}{\partial z^{(t)}}}} \left[h^{(t-1)} \right]^T$$

ii) Now look at: $\left(\frac{\partial L}{\partial h^{(t)}}\right) \boxed{\frac{\partial h^{(t)}}{\partial z^{(t)}}}$ $\mathbb{R}^n$, $\mathbb{R}^n$ as before: $\frac{\partial L}{\partial h^{(t)}_i} \boxed{\frac{\partial h^{(t)}_i}{\partial z^{(t)}_j}}$

$=1$ if $j=i$

$=0$ otherwise

$$\boxed{\frac{\partial h^{(t)}_i}{\partial z^{(t)}_j}} = \frac{\partial}{\partial z_j} \left[\tanh \left(z^{(t)}_i \right) \right] = \frac{\partial}{\partial z^{(t)}_j} \left[\boxed{\tanh} \left(\boxed{z^{(t)}_i} \right) \right] = \left[\frac{\partial}{\partial x} \tan(x) \right] \left(z^{(t)}_i \right) \left(\frac{\partial}{\partial z^{(t)}_j} z^{(t)}_i \right)$$

f, g

$$= \begin{cases} \left[\frac{\partial}{\partial x} \tan(x) \right] \left(z^{(t)}_i \right) \cdot 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

$$\boxed{\frac{d}{dx} \tanh x = 1 - (\tanh x)^2} \qquad f \quad 1 - f^2$$

$$\left(\quad \right)^2 \qquad \left[\left(1 - \left(h^{(t)}_i \right)^2 \right) \text{ if } i = j \right.$$

$\{ \; 0$ otherwise

$$= \begin{cases} 1 - (\tanh z_i^{(t)})^2 & \text{if } i=j \\ 0 & \text{otherw} \end{cases} = \begin{cases} 1 - (h_i^{(t)})^2 & \text{if } i=j \\ 0 & \text{otherw} \end{cases}$$

$$\frac{\partial L}{\partial z_j^{(t)}} = \frac{\partial L}{\partial h^{(t)}} \frac{\partial h^{(t)}}{\partial z_j^{(t)}} = \sum_i \frac{\partial L}{\partial h_i^{(t)}} \frac{\partial h_i^{(t)}}{\partial z_j^{(t)}} = \underbrace{\sum_{i \neq j} \frac{\partial L}{\partial h_i^{(t)}} \underbrace{\frac{\partial h_i^{(t)}}{\partial z_j^{(t)}}}_{=0}}_{=0} + \frac{\partial L}{\partial h_j^{(t)}} \underbrace{\frac{\partial h_j^{(t)}}{\partial z_j^{(t)}}}_{= 1 - (h_j^{(t)})^2}$$

$$= \left(1 - (h_j^{(t)})^2\right) \cdot \frac{\partial L}{\partial h_j^{(t)}}$$

$$\frac{\partial L}{\partial z^{(t)}} = \operatorname{diag}\left(1 - (h^{(t)})^2\right) \frac{\partial L}{\partial h^{(t)}}$$

$$\to \sum_t \frac{\partial L}{\partial h^{(t)}} \frac{\partial h^{(t)}}{\partial z^{(t)}} \left[h^{(t-1)}\right]^T = \sum_t \operatorname{diag}\left(1 - (h^{(t)})^2\right) \frac{\partial L}{\partial h^{(t)}} \left[h^{(t-1)}\right]^T = \frac{\partial L}{\partial W}$$

$\frac{\partial L}{\partial U}$: Is the same derivation as for $\frac{\partial L}{\partial W}$ but we interchange W and U, $h^{(t-1)}$ and $x^{(t)}$

$$= \sum_t \operatorname{diag}\left(1 - h^{(t)}\right)^2 \frac{\partial L}{\partial h^{(t)}} \left[x^{(t)}\right]^T$$

Problem 3: What do you need to change in the equations that you got in the previous exercise if the output $o^{(t)}$ is used as an input to another neural network?

So, there is f as another Layer(s) in between

$$\frac{\partial L}{\partial \theta} = \frac{\partial L}{\partial f} \frac{\partial f}{\partial \theta} =$$